Summer 2014

Design of robust spin-transfer torque magnetic random access memories for ultralow power high performance on-chip cache applications

Xuanyao Fong
Purdue University

Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Xuanyao Fong

Entitled

Design of Robust Spin-transfer Torque Magnetic Random Access Memories for Ultralow Power High Performance On-chip Cache Applications

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

KAUSHIK ROY
Chair

BYUNGHOO JUNG

MARK S. LUNDSTROM

SUPRIYO DATTA

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): KAUSHIK ROY

Approved by: V. Balakrishnan 08-01-2014
Head of the Graduate Program Date

DESIGN OF ROBUST SPIN-TRANSFER TORQUE MAGNETIC RANDOM ACCESS MEMORIES FOR ULTRALOW POWER HIGH PERFORMANCE ON-CHIP CACHE APPLICATIONS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Xuanyao Fong

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana

To family, friends, humanity, the good times and the bad.

ACKNOWLEDGMENTS

First and foremost, I sincerely thank my Ph.D. advisor, Prof. Kaushik Roy. His guidance and advice was pivotal in making this dissertation possible. The inspiration he gave not only made learning enjoyable but also set a professional benchmark for me to emulate.

I also want to thank the members of my dissertation committee: Prof. Mark Lundstrom, Prof. Supriyo Datta, and Prof. Byunghoo Jung. The invaluable lessons I learned from their courses (insights into semiconductor physics taught by Prof. Mark Lundstrom; concepts of electronic transport taught by Prof. Supriyo Datta; basic circuit design concepts taught by Prof. Byunghoo Jung) helped strengthen the foundations that made this dissertation possible.

Next, I want to acknowledge the sponsors of this research without whom this dissertation would not have been possible: Intel Corp, Advanced Micro Devices, Inc., Qualcomm. This work was also supported in part by C-SPIN, one of six centers of STARnet, a Semiconductor Research Corporation program, sponsored by MARCO and DARPA.

I would also like to thank Prof. Arijit Raychowdhury (Georgia Institute of Technology), who mentored me during my early years as a researcher. It was his guidance and encouragement that nurtured by passion for research and persuaded me to pursue this Ph.D.

I want to thank former members of the Nanoelectronics Research Laboratory (NRL, Purdue University) for their inspiration and support: Dr. Jaydeep Kulkarni (Intel Corp.), Prof. Saibal Mukhopadhyay (Georgia Institute of Technology), Dr. Qikai Chen (Intel Corp.), Dr. Myeong-Eun Hwang (Samsung Electronics), Prof. Ik Joon Chang (Kyung Hee University), Dr. Sang Phill Park (Intel), Dr. Mesut Meterelliyoz (Intel Corp.), Dr. Patrick Ndai (Texas Instruments), Dr. Ashish Goel

(Broadcom), Dr. Jing Li (IBM), Dr. Georgios Karakonstantis (EPFL), Prof. Swaroop Ghosh (University of South Florida), Dr. Chih-Hsiang Ho (Qualcomm), and Dr. Nilanjan Banerjee (Qualcomm).

I also want to acknowledge my co-authors and collaborators for the assistance as well as the knowledge I gained from my experience with them: Prof. Sumeet Gupta (Penn State University), Dr. Rangharajan Venkatesan (Nvidia Corp.), Dr. Charles Augustine (Intel Corp.), Dr. Niladri Mojumder (Qualcomm), Dr. Sri Harsha Choday (Qualcomm), Dr. Georgios Panagopoulos (Intel Corp.), Dr. Behtash Behin-Aein (Globalfoundries), Dr. Dongsoo Lee (IBM) and Dr. Mrigank Sharad.

I would like to thank other members of NRL for their support: Yusung Kim, Kon-woo Kwon, Deliang Fan, Karthik Yogendra, Mei-chin Chen, and others I might have accidentally missed out.

Next, I want to thank my friends for their support and encouragement throughout my graduate studies, as well as the life lessons they taught me: Dr. Amanda Lee and Ian Tan, Dr. Scott Poh, Dr. Wing Fai Loke and Wendy Woon, Dr. Shisheng Huang and Mun Yee Tham, Joshua Chia and Linnet Foong, Prof. KwekTze Tan (University of Akron), Prof. Nelson Wei Tan (San Jose State University), Jignesh Mehta, Zherui Guo, Dinesh Sandran, Ying Zhi Pao, Praveen Kumar, Tiffany Sukwanto, Dr. Winnie Tan, Melanie Wong, Armando Indrajuana, Peter Adjiwibawa, and many others that I am able to list here.

Finally, I want to thank my parents and my family for the love, support and encouragement they provided throughout this dissertation. I am forever indebted to them.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Fong, Xuanyao Ph.D., Purdue University, December 2014. Design of Robust Spin-transfer Torque Magnetic Random Access Memories for Ultralow Power High Performance On-chip Cache Applications. Major Professor: Kaushik Roy.

Spin-transfer torque magnetic random access memories (STT-MRAMs) based on magnetic tunnel junction (MTJ) has become the leading candidate for future universal memory technology due to its potential for low power, non-volatile, high speed and extremely good endurance. However, conflicting read and write requirements exist in STT-MRAM technology because the current path during read and write operations are the same. Read and write failures of STT-MRAMs are degraded further under process variations. The focus of this dissertation is to optimize the yield of STT-MRAMs under process variations by employing device-circuit-architecture co-design techniques. A devices-to-systems simulation framework was developed to evaluate the effectiveness of the techniques proposed in this dissertation. An optimization methodology for minimizing the failure probability of 1T-1MTJ STT-MRAM bit-cell by proper selection of bit-cell configuration and access transistor sizing is also proposed. A failure mitigation technique using *assists* in 1T-1MTJ STT-MRAM bit-cells is also proposed and discussed. Assist techniques proposed in this dissertation to mitigate write failures either increase the amount of current available to switch the MTJ during write or decrease the required current to switch the MTJ. These techniques achieve significant reduction in bit-cell area and write power with minimal impact on bit-cell failure probability and read power. However, the proposed write assist techniques may be less effective in scaled STT-MRAM bit-cells. Furthermore, read failures need to be overcome and hence, read assist techniques are required. It has been experimentally demonstrated that a class of materials called

multiferroics can enable manipulation of magnetization using electric fields via magnetoelectric effects. A read assist technique using an MTJ structure incorporating multiferroic materials is proposed and analyzed. It was found that it is very difficult to overcome the fundamental design issues with 1T-1MTJ STT-MRAM due to the two-terminal nature of the MTJ. Hence, multi-terminal MTJ structures consisting of complementary polarized pinned layers are proposed. Analysis of the proposed MTJ structures shows significant improvement in bit-cell failures. Finally, this dissertation explores two system-level applications enabled by STT-MRAMs, and shows that device-circuit-architecture co-design of STT-MRAMs is required to fully exploit its benefits.

1. INTRODUCTION

The evolution of the semiconductor industry over the past few decades has been driven mainly by technology scaling. The density of on-chip transistors increased significantly from ∼700 transistors per mm^2 in 1980 to ∼6 million transistors per mm^2 in state-of-the-art microprocessors today as shown in Fig. 1.1 [1–5]. Benefits of technology scaling in microprocessors include increased functionality and more than 100× performance increase. The performance increase is due to faster switching speed of the transistors as well as increased on-chip cache size, which reduces the number of cache misses that significantly impact the throughput of the processor [7]. However, frequency scaling has been limited by the significant increase in power dissipation density due to technology scaling [1, 7] and processor operating frequencies have reached a plateau recently, as shown in Fig. 1.2 [8, 9]. The power densities in state-of-the-art processors today are ∼65W/cm^2 [2] and is approaching that in nuclear reactors if nothing is done. Designers have attempted to mitigate the power dissipation problem using architectural techniques such as multi-cores (the trend of core counts is shown in Fig. 1.3) and increased on-chip cache size. Radical changes in software develop-

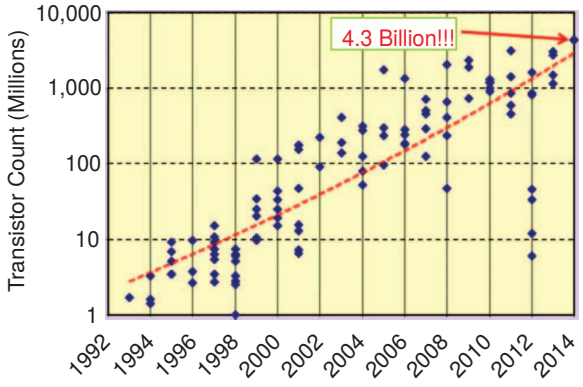


Fig. 1.1. CMOS scaling trends in terms of transistor count as reproduced from [6].

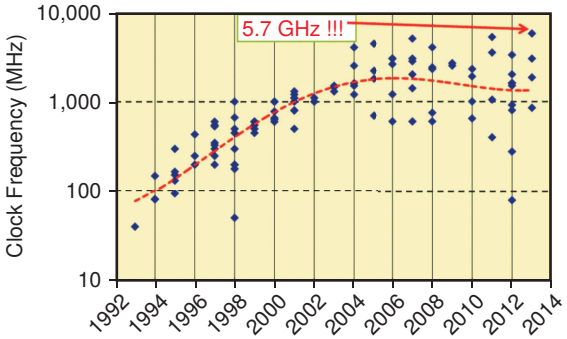


Fig. 1.2. CMOS scaling trends in terms of operating frequency as reproduced from [9].

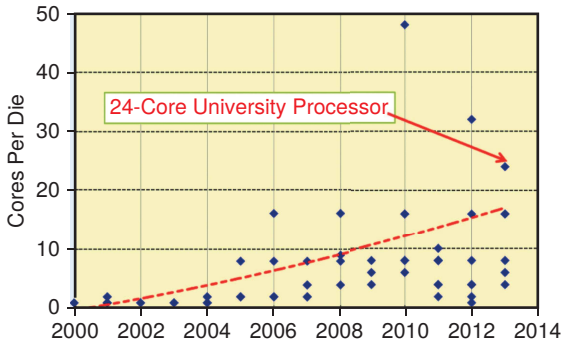


Fig. 1.3. CMOS scaling trends in terms of core count as reported from [9].

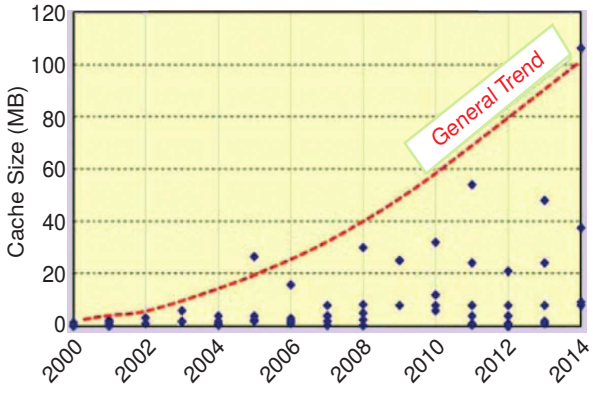


Fig. 1.4. Scaling trend of on-chip caches reported in [6].

ment are needed to take full advantage of multi-cores [10]. Alternatively, smaller transistor size allows chip designers to increase the size of on-chip caches (Fig. 1.4) to enhance chip performance by keeping as much data as possible close to the processor cores. State-of-the-art 6T SRAM may occupy as much as ∼40% of core area in microprocessors today [2].Thus, power dissipation in modern microprocessors is increasingly dominated by leakage power in the memory subsystems. Furthermore, memory subsystems based on the 6T SRAM cells lose its stored data when turned off so they cannot be turned off to save power when the processor is idle [1]. Hence, low power and high speed non-volatile memory technologies compatible with current CMOS technology is needed to mitigate the huge power dissipation due to technology scaling while increasing cache size.

Several non-volatile memory technologies have emerged and are intensively researched recently [11]. The most attractive memory technologies that have been identified are phase-change memory (PCM), ferroelectric RAM (FeRAM), magnetic RAM (MRAM) and resistive RAM (RRAM). However, MRAM has emerged as the leading candidate for future universal memory because of its potential for high-performance (< 10 ns) and extremely good endurance ($> 10^{14}$ write cycles) compared to other non-volatile memory technologies. MRAMs are also compatible with the CMOS fabrication process, requiring minimal changes to the back-end-of-line (BEOL) fabrication process (addition of 2 mask steps). The basic storage device in MRAMs is the magnetic tunnel junction (MTJ). MRAMs based on the MTJ are inherently compatible with digital logic because the MTJ have only two stable resistance states [12]. The basics of and design issues in MRAMs will be discussed in the following sections.

1.1 The Magnetic Tunnel Junction

The magnetic tunneling junction (MTJ) is used as the storage element in MRAMs. The structure of an MTJ is illustrated in Fig. 1.5(a) and it consists of a tunneling oxide layer (MgO has replaced Al_2O_3 as the tunneling oxide because its crystalline structure

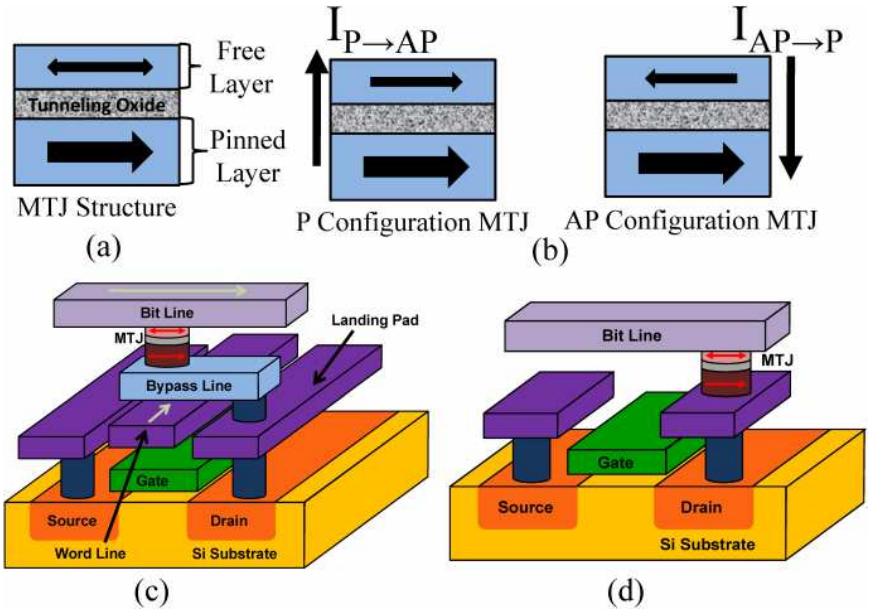


Fig. 1.5. (a) Structure of a magnetic tunnel junction, (b) Charge current directions to induce spin-transfer torque switching, (c) bit-cell structure of field-switched MRAM and of (d) spin-transfer torque MRAM (STT-MRAM)

enhances the tunneling magnetoresistance ratio of the MTJ, which will be discussed later) sandwiched between two ferromagnetic electrodes. One of the ferromagnetic electrodes is magnetically pinned (called the pinned layer or PL) so that it can be used as a reference layer. The other ferromagnetic electrode (called the free layer or FL) is engineered so that its magnetization direction can be either parallel (P) or anti-parallel (AP) to that of the PL. The energy barrier between P and AP configurations is small enough such that the MTJ can be switched between configurations but large enough to ensure thermal stability. The electrical characterization of individual MTJs has been reported in [13–16]. Binary data is represented by and stored as the magnetic configuration of an MTJ, which may be sensed as the MTJ resistance, R_{MTJ}, as will be discussed later.

A metric for the MTJ as shown in [13] is its resistance-area (RA) product. At iso-cross-sectional area, R_{MTJ} depends exponentially on the tunneling oxide thickness,

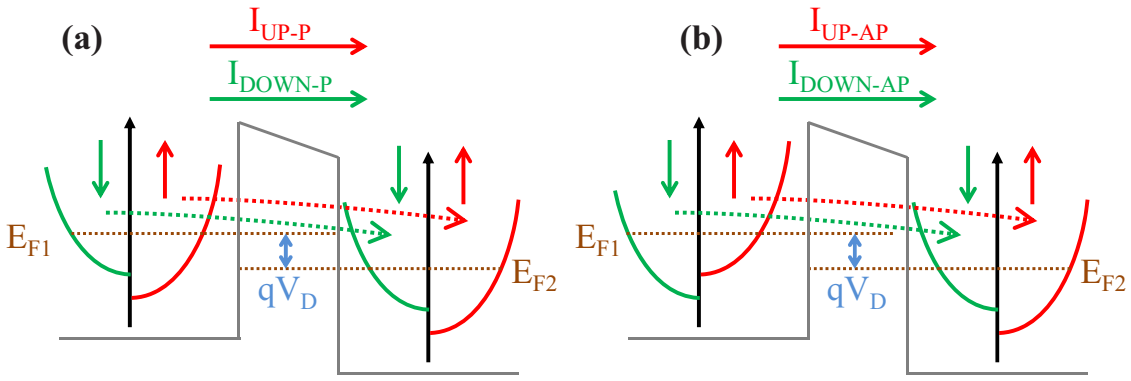


Fig. 1.6. Band diagrams for up and down spins when MTJ is in (a) parallel configuration and in (b) anti-parallel configuration, to illustrate effect of tunneling magneto-resistance. Parabolic bands depict the lowest conduction band in the magnetic layers.

t_{MgO}, since the mechanism for electron transport is direct tunneling. At iso-t_{MgO}, R_{MTJ} depends linearly on the cross-sectional area of the MTJ, A_{MTJ}, similar to an Ohmic conductor. R_{MTJ} also depends on the relative magnetization direction of the FL with respect to the PL, which is also called the *tunneling magneto-resistance effect*. The tunneling magneto-resistance effect arises due to the difference in density of states around the Fermi energy (E_F) of the ferromagnetic contacts [17].

Fig. 1.6 illustrates an example band structure of the MTJ when it is in (a) the P configuration and in (b) the AP configuration. Electrons flowing between the electrodes carry either up-spin (majority spin) or down-spin (minority spin). Assuming that spin scattering is negligible, the flow of minority and majority spins can be thought of as two decoupled current paths (I_{MAJ} or I_{UP} for majority spins, and I_{MIN} or I_{DOWN} for minority spins) and the total charge current flow is $I_{UP} + I_{DOWN}$.

Consider the MTJ in the P configuration first. Fig. 1.6(a) illustrates the band diagram along the electron transport direction of an MTJ in the P configuration. The density of states for like-spins in the FL matches that in the PL and when a small voltage, V_D, is applied, there are sufficient states to accommodate all the electrons available for conduction. Note that $I_{UP-P} > I_{DOWN-P}$ since the density of states for

up-spin electrons is higher than that of down-spins in both PL and FL. Furthermore, the total charge current, I_{CH-P}, obeys the inequality $I_{CH-P} > 2I_{DOWN-P}$.

Now consider the MTJ in AP configuration instead. Fig. 1.6(b) illustrates the band diagram along the electron transport direction of an MTJ in the AP configuration. Further consider when a small voltage, V_D, is applied such that the bands in the left electrode are raised relative to the bands in the electrode on the right. Note that there is a mismatch between density of states of like-spins in the electrodes on left and on the right. There are more down-spin electrons available for conduction in the left electrode than the number of available down-spin states to fill in the right electrode. On the other hand, there are more up-spin states available for conduction in the right electrode than number of up-spin electrons available for conduction in the left electrode. Hence, I_{UP-AP} is limited by the number of up-spin electrons available in the left electrode while $I_{DOWN-AP}$ is limited by the number of down-spin states available in the right electrode for conduction. Note that $I_{UP-AP} \approx I_{DOWN-AP}$ because the number of electrons that can flow between the electrodes are the about the same for up-spins and for down-spins.

Note that $I_{DOWN-P} \approx I_{DOWN-AP} \approx I_{UP-AP}$. Thus, the charge current, I_{CH-AP}, that flows when V_D is applied across the MTJ in the AP configuration is such that $I_{CH-AP} \approx 2I_{DOWN-AP} \approx 2I_{DOWN-P} < I_{CH-P}$. Hence, for the same applied voltage, V_D, more charge current flows across the MTJ when it is in the P configuration than when it is in the AP configuration. Thus, R_{MTJ} is low in the P configuration ($R_P = R_L$) and high in the AP configuration ($R_{AP} = R_H$). The *Tunneling Magneto-resistance Ratio* $\left(TMR = 100\% \times \frac{R_H - R_L}{R_L}\right)$ measures the difference in MTJ resistance between P and AP configurations and is another metric for the performance of MTJs as a resistive memory element.

1.2 MRAM Read and Write Operations

The bit-cell of field-switched MRAM is shown in Fig. 1.5(c). The MTJ state is switched using magnetic fields generated from current-carrying word and bit lines. However, field-switched MRAMs are not scalable for two reasons [12]. First, the magnetic fields used for switching the MTJ are not confined to individual bit-cells and may cause unintended writing into neighboring cells in very dense field-switched MRAM arrays. Second, the current required to generate the magnetic field for writing increases with scaling. When the MTJ size is scaled down, the critical field needed to switch the MTJ needs to be scaled up proportionally to maintain thermal stability and retention time. The retention time, t_{RET}, depends on the energy barrier of the free layer in the MTJ and is given by [12]

$$t_{RET} = t_0 e^{\frac{E_A}{k_B T}} \tag{1.1}$$

where t_0 is on the order of 1 ns, E_A is the activation energy or energy barrier of the free layer, k_B is the Boltzmann constant and T is the temperature in Kelvin. For a single-bit, the energy barrier needs to be $\sim 40 k_B T$ for 10 years of retention time. The anisotropy energies in the free layer of the MTJ are engineered to achieve the energy barrier required to achieve the desired retention time. In the simplest form, the uniaxial anisotropy energy (discussed in Appendix B) is engineered to achieve the required retention time. The critical magnetic field needed to switch the MTJ configuration is calculated from the activation energy by [12]

$$H_C = \frac{2E_A}{\mu_0 M_S V} \tag{1.2}$$

where μ_0 is the permeability of free space, M_S is the saturation magnetization of the free layer material, and V is the volume of the free layer in the MTJ. Consider if the cross-sectional area of the MTJ is scaled down by some factor, κ, to keep pace with the scaling down of CMOS technology. If the thickness of the free layer in the MTJ is kept constant, the critical field of the MTJ needs to be scaled up by κ to maintain the retention time of the MTJ. Hence, the amount of current needed to

generate sufficient magnetic field to program the MTJ increases with the scaling down of MTJ size. Such inherent scaling issues with field-switching have led researchers to investigate alternate methods for magnetization reversal in MTJs.

There are two flavors of MTJs available: those with ferromagnetic layers having *in-plane magnetic anisotropy* (IMA) and those with ferromagnetic layers having *perpendicular magnetic anisotropy* (PMA). The easy magnetization direction (explained in Appendix B) of a ferromagnetic layer with IMA lies within the plane of the thin film ferromagnetic layer. On the other hand, the easy magnetization direction of a ferromagnetic layer with PMA lies perpendicular to the plane of the thin film ferromagnetic layer. The demagnetizing field of a thin film ferromagnet tends to align in the direction perpendicular to the plane of the thin film ferromagnetic layer. If the ferromagnetic layers are engineered with uniaxial anisotropy to achieve the desired retention time, the demagnetizing field will be perpendicular to the uniaxial anisotropy field in the ferromagnetic layer with IMA, whereas the two fields are collinear in the ferromagnetic layer with PMA. As will be shown in Chapter 2, the fields cancel each other in the PMA ferromagnet leading to a lower switching field needed to switch the PMA ferromagnet as compared to an IMA ferromagnet with the same retention time. A lower switching field is preferred to reduce the energy dissipation in MRAMs.

Since the prediction of spin-polarized current induced magnetization reversal by Slonczewski [18] and Berger [19], spin-transfer torque magnetic RAM (STT-MRAM) have been proposed as the solution to the non-scalability of field-switched MRAM [17]. Magnetization reversal in STT-MRAM occurs due to spin-flip processes when current flows through the MTJ perpendicular to the magnet-oxide-magnet interfaces [see Fig. 1.5(b)] and thus, is well confined within each bit-cell. Recently, STT-MRAM arrays have been fabricated and characterized [20–22].

The read operation in both field-switched MRAM and STT-MRAM are similar. The data stored in the MTJ is determined by sensing its resistance using either a voltage sensing scheme or a current sensing scheme. In the voltage sensing scheme, a fixed current is passed through the MTJ and the voltage developed across it is com-

pared to a reference voltage to determine the MTJ resistance—the voltage developed across the MTJ is lower than the reference voltage when the MTJ is in P configuration and higher than the reference voltage when the MTJ is in AP configuration. Alternatively, a fixed voltage is applied across the MTJ in the current sensing scheme. The current flowing through the MTJ is then compared to a reference current—the MTJ current is higher than the reference current when the MTJ is in the P configuration and lower than the reference current when the MTJ is in the AP configuration. The advantages and disadvantages of these sensing schemes will be discussed later.

1.3 Design of Spin-Transfer Torque MRAM Bit-cell

Several STT-MRAM bit-cell designs have been published in the literature [20–22]. As shown in Fig. 1.5(d), the 1T-1MTJ (or 1T-1R) STT-MRAM bit-cell stores a single-bit of data and consists of an NMOS transistor (NFET) and an MTJ. The word line turns the NFET on or off. When the NFET is on, charge current can flow through the MTJ when there is a voltage difference between bit line (BL) and source line (SL). Depending on the magnitude and direction of the current, the MTJ configuration may be manipulated by spin-transfer torque as predicted by Slonczewski [18] and Berger [19]. In on-chip memory applications, small bit-cell areas are preferred so that as much data as possible may be stored in a fixed area on the silicon die (measured using a metric called *memory density*). Thus, this discussion focuses on the 1T-1R STT-MRAM bit-cell. According to the ITRS roadmap [23], the bit-cell area of 1T-1R STT-MRAM is expected to be dominated by the NFET size. However, the NFET size depends on the electrical resistance of the MTJ. 1T-1R STT-MRAM bit-cells can have two configurations [20, 21] as shown in Fig. 1.7: the "standard" connection [SC, Fig. 1.7(a)] and the "reversed" connection [RC, Fig. 1.7(b)]. An objective of this dissertation is to establish which connection has better yield under process variations.

Magnetization reversal in the FL of the MTJ occurs when the current density flowing through the MTJ exceeds a threshold value [12, 24] (also known as the *critical*

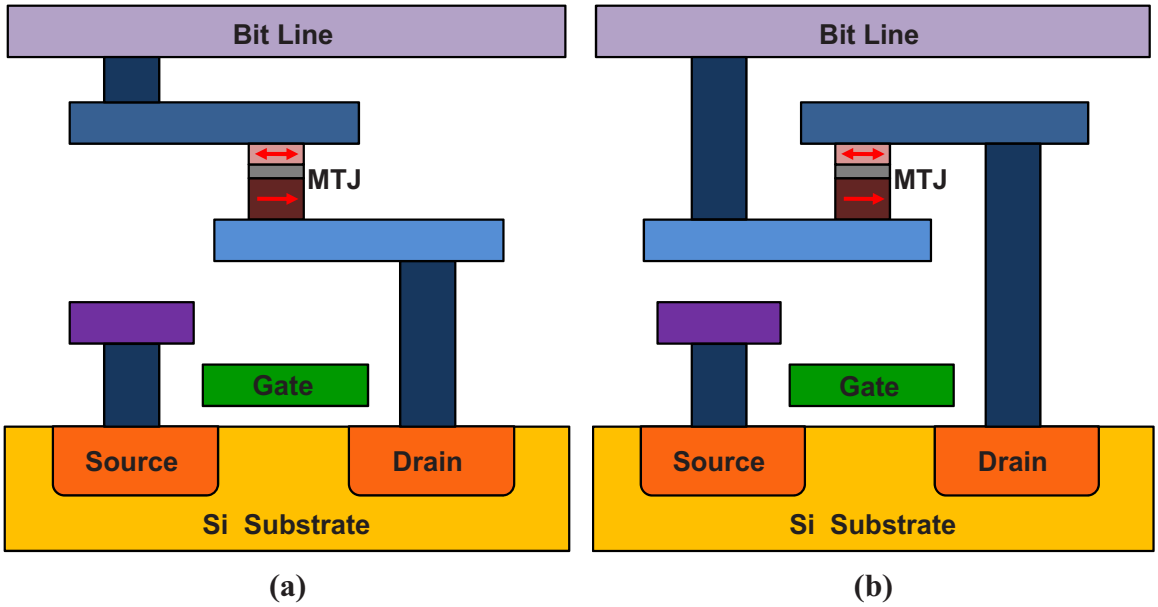


Fig. 1.7. Structures of a (a) "standard" connection or SC 1T-1MTJ STT-MRAM bit-cell and a (b) "reversed" connected or RC 1T-1MTJ STT-MRAM bit-cell.

current density, J_C). However, an inherent asymmetry in J_C exists in switching an MTJ from the P configuration to the AP configuration compared to switching from the AP configuration to the P configuration. In an MTJ, the PL acts as a spin filter that polarizes the tunneling current. When electrons flow from the PL to the FL, the electrons are first spin polarized by the PL before tunneling across the tunneling oxide into the FL. Most of the electrons entering the FL are spin polarized in the direction of the magnetization of the PL, and they exert a spin-transfer torque on the FL to orient the FL magnetization parallel to that of the PL. When electrons flow from the FL to the PL, the electrons entering the FL are not spin polarized and may have any spin direction. Since the FL is also a ferromagnet, it tries to polarize the spin of the incoming electrons with its magnetization direction. However, electrons with the same spin polarization as the PL magnetization direction may tunnel easily across the tunneling oxide and hence, are easily removed from the FL. During P to

AP switching of the MTJ, electrons with spin polarization opposite the magnetization direction of the PL exchange spin angular momentum with the FL in order to become spin polarized in the direction of PL magnetization. Hence, these electrons exert a spin-transfer torque to align the FL magnetization opposite that of the PL before they are easily removed from the FL. Since there are much fewer electrons exerting spin-transfer torque to switch the FL magnetization anti-parallel to that of the PL than to switch the FL magnetization parallel to that of the PL, it appears as though the spin polarization efficiency depends on the direction of current flow through the MTJ. The spin polarization efficiency is high when electrons are flowing from the PL into the FL, and low when the electrons are flowing from the FL into the PL. Hence, there is an asymmetry in critical switching current densities [25, 26]. It has been reported that the J_C when switching the MTJ from P to AP configuration can be 10% to 2× larger than when switching from AP to P configuration [21, 27–29]. This may be an important design issue as explained later.

1.4 Design Issues in STT-MRAM

The fundamental improvements desired in STT-MRAM are in its 1) read performance, 2) write performance, 3) retention time and thermal stability, and 4) reliability. However, it is extremely challenging to achieve these improvements simultaneously in STT-MRAM due to conflicting design requirements. For example, J_C is increased if the thermal stability of the FL is increased, as the later sections will show. Many of the conflicting design requirements in STT-MRAM occur because of three fundamental design issues: *source degeneration of ATx during write operations*, *shared read and write current paths*, and *single-ended sensing of stored data*.

A severe design issue arising from the need for bi-directional write current in STT-MRAM is that the NFET is source degenerated when current flows from the source-line to the bit-line during write operations. Consider the voltage biases in the bit-cell as shown in Fig. 1.8. When current is flowing from bit-line (BL) to source-line

(SL), the source of the NFET is the terminal connected to SL. Hence, the bias on the NFET is such that $V_{GS} = V_{DD}$. When current flows from SL to BL instead, the source of NFET at the terminal connected to the MTJ. Denoting the voltage on the source terminal of the NFET as V_X, Fig. 1.8 shows that $GND < V_X < V_{DD}$. Hence, the bias on the NFET is such that $V_{GS} < V_{DD}$. This means that the NFET size may need to be increased to allow sufficient current to flow from SL to BL. Increasing the width of the NFET also leads to an increase in the current flowing from BL to SL during write operations, which may be excessive and may lead to excessive write power dissipation and degradation in the reliability of the tunnel oxide in the MTJ. The reliability of the tunnel oxide is crucial to maintain the TMR of the MTJ and hence, the distinguishability of the MTJ states.

As mentioned in Section 1.2, a voltage or current sensing scheme is used to sense the resistance of the STT-MRAM bit-cell during read operations. Regardless of the scheme used, a current flows through the MTJ during read operations. Fig. 1.9 illustrates an example biasing condition for reading data stored in STT-MRAM bit-cells using the current sensing scheme. Note that the current flowing through the bit-cell during read operations may also be increased if the NFET width is increased. If the read current is sufficiently large, the MTJ may be accidentally overwritten during

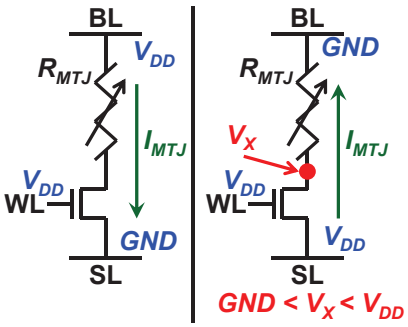


Fig. 1.8. The voltages in the STT-MRAM bit-cell when (left) current flows from bit-line, BL, to source-line, SL, and when (right) current flows from SL to BL. The word-line, WL, switches the access transistor on and off.

Table 1.1.
STT-MRAM control voltages during read and write operations

$V_{WL} = V_{DD}$	SC	RC
Write (Parallelizing)	$V_{BL} = V_{DD}$ $V_{SL} = GND$ $I_{MTJ} \geq I_C$ (AP $\rightarrow$ P)	$V_{BL} = GND$ $V_{SL} = V_{DD}$ $I_{MTJ} \geq I_C$ (AP $\rightarrow$ P)
Write (Anti-parallelizing)	$V_{BL} = GND$ $V_{SL} = V_{DD}$ $I_{MTJ} \geq I_C$ (P $\rightarrow$ AP)	$V_{BL} = V_{DD}$ $V_{SL} = GND$ $I_{MTJ} \geq I_C$ (P $\rightarrow$ AP)
Read (Parallelizing)	$V_{BL} = V_{READ} < V_{DD}$ $V_{SL} = GND$ $I_{MTJ} < I_C$ (AP $\rightarrow$ P)	$V_{BL} = GND$ $V_{SL} = V_{READ} < V_{DD}$ $I_{MTJ} < I_C$ (AP $\rightarrow$ P)
Read (Anti-parallelizing)	$V_{BL} = GND$ $V_{SL} = V_{READ} < V_{DD}$ $I_{MTJ} < I_C$ (P $\rightarrow$ AP)	$V_{BL} = V_{READ} < V_{DD}$ $V_{SL} = GND$ $I_{MTJ} < I_C$ (P $\rightarrow$ AP)

read operations because the read and write current paths are shared (also known as *read-disturb failure*, which will be discussed further in Section 3.1.2). Table 1.1 shows the voltages on the control lines of the bit-cell and the current flowing through it

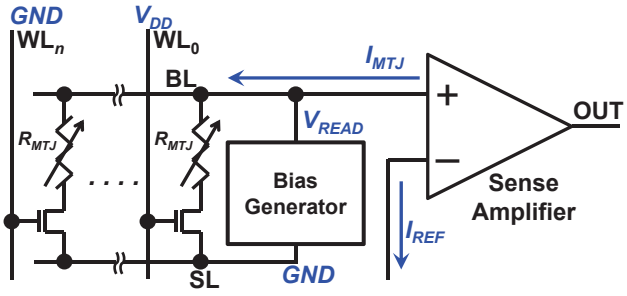


Fig. 1.9. The circuit description of the sensing scheme for performing read operations in one column of the STT-MRAM array, which consists of n rows. Only Row 0 is selected and all other rows are not ($V_{WL} = 0$ V for unselected cells). The voltages on the control lines indicate one possible configuration for sensing. The direction of I_{MTJ} is reversed if the voltages on BL and SL are swapped and hence, the direction of I_{REF} needs to be swapped too.

(for "SC" and "RC" bit-cell configurations). The amount of current flowing through the bit-cell during read operations needs to be limited to avoid disturbing the bit-cell during read operations, and doing so may degrade the performance of read operations. If the amount of read current is too small, the sense amplifier may not be able to distinguish the state of the MTJ. Also, since the read current charges up the internal and input capacitances of the sense amplifier, a reduced read current means that it will take longer for the voltages on these capacitances to stabilize. Hence, the sensing delay may be increased as well if the read current is limited. The advantage of the voltage sensing scheme is that the current flowing through the bit-cell may be effectively limited, and the data stored in the bit-cell is sensed as a voltage. However, the MTJ resistance decreases with the voltage across the MTJ. Since the voltage across the MTJ in AP configuration is larger than when the MTJ is in P configuration, the TMR of the MTJ is reduced when voltage sensing scheme is used, resulting in degraded distinguishability of the stored MTJ states. On the other hand, the current sensing scheme allows the TMR of the MTJ to be fixed by clamping the voltage across the MTJ, thus improving the distinguishability of MTJ states. However, the stored MTJ data is sensed as a current and will need to be converted to a voltage before it may be used by other circuits. Furthermore, it may not be easy to control the current flowing through the MTJ in the current sensing scheme—it needs to be limited to avoid read-disturb failure but also sufficiently large for the sense amplifier to be able to distinguish the stored MTJ state.

The read-disturb failure just described may be avoided if the read operations are sufficiently fast. As shown in [12] and later in this dissertation, there is an increase in J_C and hence I_C, when the target switching delay is reduced. Thus, when the read delay is small, large read currents may be tolerated before the MTJ is accidentally overwritten. However, the single-ended nature of the sensing operation in STT-MRAM may limit the achievable read speed. Furthermore, single-ended sensing schemes without self-referencing are more prone to failures under process variations. Consider a single-ended sensing scheme where the voltage across BL and SL are fixed,

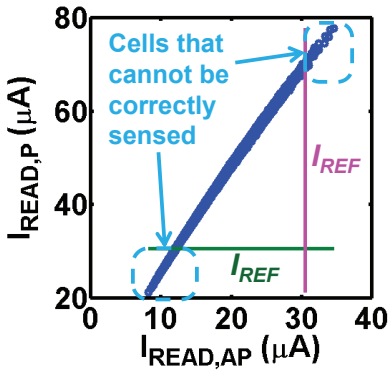


Fig. 1.10. This scatter plot conceptually describes single-ended sensing in STT-MRAM, showing that it is prone to sensing errors under process variations.

the ATx is turned on and the current flowing through the bit-cell is compared to a global reference current, I_{REF}. Fig. 1.10 shows a scatter plot of the read currents flowing through a bit-cell, which is generated using the model that will be presented in Chapter 2. Each point on the scatter plot corresponds to a bit-cell in which the read current, I_P, flows through the bit-cell when its MTJ is in P configuration, and I_{AP} flows through the same bit-cell when its MTJ is in the AP configuration. Using the single-ended sensing scheme described earlier, all bit-cells falling to the right of the vertical line will be always be sensed as P, whereas those falling below the horizontal line will always be sensed as AP. Note that there is a strong correlation between I_{AP} and I_P. A self-referencing scheme is required to exploit the correlation between I_{AP} and I_P to reduce sensing failures. Self-referencing schemes were proposed in [30, 31]. However, the proposed sensing schemes require several read operations to achieve self-referencing. Thus, a self-referenced differential sensing scheme, in which the data is stored as a pair of complementary values, is needed to improve the read performance of STT-MRAM. Since the sensing scheme is differential in nature, it is unaffected by process variations that skew the characteristics of the complementary values in the same direction.

The various device-circuit-system co-design techniques developed in this dissertation to overcome the aforementioned design issues in STT-MRAM will be presented in the following chapters. A survey of the literature is presented next to discuss the advantages and disadvantages of some of the previously proposed STT-MRAM design techniques. It should be emphasized that the design techniques proposed in this dissertation complement existing design techniques so as to fully realize the true potential of STT-MRAM.

1.5 Prior Art on Device-Circuit-Architecture Co-design of STT-MRAMs

1.5.1 Modeling of the magnetic tunnel junction

Several models for MTJs have been previously proposed [32–35]. Many of these models are simple compact models and may not capture all the necessary physical phenomena in the MTJ (such as the magnetization dynamics in the FL of the MTJ). On the other hand, micromagnetic models are used separately to simulate the dynamics of the FL in the MTJ and to estimate MTJ performance [36]. These simulations do not include effects due to electron transport in the MTJ and due to channel resistance of the access transistor. Hence, a model that captures all the physics in an MTJ is needed to evaluate effectiveness of optimization and failure mitigation techniques for STT-MRAM bit-cells.

The models proposed in [32, 33] do not capture the stochastic nature of MTJ switching. The stochastic nature of MTJ switching is an important effect because bit-cell requirements for writing are usually associated with a corresponding *write error rate* (WER) [23]. The proposed models are more suitable for modeling MTJ write operations in the precessional regime [12]. The write currents in the *precessional regime* are very large and can cause reliability issues in the MTJ. Hence, MTJ writes are usually done in the dynamic and thermal regimes. The stochastic nature of MTJ switching needs to be captured when simulating MTJ write operations in these regimes.

In the model proposed in [34, 35], magnetization dynamics in the MTJ are not modeled. The proposed model is compatible with the HSPICE circuit simulator [37] but uses a stochastic block to model the stochastic nature of MTJ switching. As will be shown in Chapter 2, this model does not capture the correlation between switching events and may not accurately predict certain failure mechanisms.

The MTJ model proposed in [35] does not include accurate simulation of electron transport in the MTJ and hence, requires the MTJ to be fabricated and characterized to calibrate the model before simulations. Such an approach is not cost effective and does not allow STT-MRAM designers to investigate the impact of material choice and parameters on the design space of STT-MRAM.

An objective of this dissertation is to propose optimization and failure mitigation techniques for developing robust STT-MRAM bit-cells and hence, an accurate MTJ model is required to develop and evaluate our proposed optimization and failure mitigation techniques. Thus, an MTJ model that captures stochastic effects due to non-zero temperature, magnetization dynamics, and atomistic electron transport in the MTJ was developed as part of this dissertation. The proposed model is then used to predict variations in electrical characteristics of MTJs due to process variations. This approach is more accurate because no assumptions are made about electrical characteristics of MTJs. Details of the proposed MTJ model will be discussed in Chapter 2.

1.5.2 Architecture-level STT-MRAM design techniques

Architectural techniques to design robust STT-MRAM arrays have been proposed in the literature [38–41]. The stretched write cycle (SWC) technique described in [38,39] exploits the fact that memory writes occur much less frequently than memory read operations to mitigate write failures. Since the required current to program an STT-MRAM bit-cell reduces with operating frequency, SWC allows more time for writing into STT-MRAM and hence, reduces the write failure probability.

Alternatively, redundancy techniques such as those proposed in [40, 41] may be used to mask bit-cell failures in an array. The small bit-cell area achievable in STT-MRAM allows more bit-cells to be packed into an array compared to SRAMs. At the architecture-level, the memory capacity required may be lesser than the number of bit-cells in the memory array (these bit-cells are *redundant cells*). When bit-cells that do not function properly are detected, the data is written to or taken from some other corresponding bit-cell instead. The remapping of bit-cells may be done prior to chip operation or during chip operation. Since the number of faulty bit-cells tolerable in the array increases with the number of redundant bit-cells, the yield of the memory array is improved if more redundant bit-cells are available to mask faulty bit-cells.

1.5.3 Circuit-level STT-MRAM design techniques

Circuit-level design techniques for robust STT-MRAM have also been proposed in the literature [42, 43]. The MTJs that were used in the analysis performed in [42, 43] are very prone to read disturb failures. In the 1T-1MTJ STT-MRAM, electrical current flows through the MTJ during both read and write operations. When the access transistor is sized up to reduce write failures, more current can flow through the MTJ during read operations and read-disturb failures are increased. The technique proposed in [42, 43] uses a 2T-1MTJ STT-MRAM bit-cell topology consisting of two access transistors instead of one. Both access transistors are turned on during write operations to maximize write current flowing through the MTJ. Only one access transistor is turned on during read operation to limit the current flowing through the bit-cell. However, doing so may degrade the bit-cell TMR and make it more difficult to sense the data stored in the bit-cell.

1.5.4 Device-level STT-MRAM design techniques

Alternative MTJ structures have been proposed to mitigate the conflicting design requirements for read and for write, which are inherent in the conventional MTJ

structure [44–47]. The read and the write current paths are decoupled in these devices, even though data is stored in a common free layer. Furthermore, they have separate read and write ports that allow independent optimization for read and for write. Another advantage of these structures is that the current tunneling through the oxide used for read operations is always limited, which improves the oxide reliability and hence, the lifetime of the bit-cell. However, the scalability of the MTJs and the integration density of the bit-cell using them may be degraded because they need more than one access transistor.

The work done by the research community on STT-MRAM bit-cell device-circuit-architecture co-design was reviewed in the preceding sections. An important observation is that models used in many of these analyzes may not be accurate enough. An accurate model that is compatible for similar analysis and optimization of STT-MRAM is developed as part of this dissertation. Furthermore, the failure mitigation techniques developed in this dissertation complement the techniques that have been proposed in the literature. A significant contribution of this dissertation is the addition of new design techniques that may be used concurrently with other existing techniques to design robust high performance and high density on-chip STT-MRAM.

1.6 Summary

This chapter reviewed the fundamentals of STT-MRAMs that are necessary for motivating and understanding the work presented in the later chapters of this dissertation. As explained earlier in this chapter, it is desirable to improve several aspects of the design and operation of the standard STT-MRAM: write-ability, readability, thermal stability, and reliability. It remains challenging to do so because these metrics have conflicting design requirements, which will be a recurring theme in the rest of this dissertation. A survey of the literature on device-circuit-architecture co-design of STT-MRAMs was also presented to discuss the failure mitigation techniques that have been proposed to improve STT-MRAMs. However, STT-MRAM models, op-

timization and failure analysis methodologies are needed to evaluate the efficacy of the proposed failure mitigation techniques. Hence, the models and methods used in the prior art are discussed in this chapter along with their shortcomings. An important observation is that models used in many of these analyzes may not be accurate enough. An accurate model that is compatible for similar analysis and optimization of STT-MRAM is developed as part of this dissertation and will be presented in the next chapter. After presenting the modeling and simulation framework used in this dissertation, the focus of the discussion will shift to the failure analysis methodology developed in this dissertation and the STT-MRAM design techniques proposed to overcome and mitigate design issues in STT-MRAM. A significant contribution of this dissertation is that the models, failure analysis methodology, and design techniques proposed in this dissertation complements existing work on the design of STT-MRAM in the literature.

The rest of this dissertation is organized as follows. The fundamentals and issues in STT-MRAMs have been discussed in the preceding sections of this chapter. The research work already done by the research community to address these issues was also discussed. However, there are shortcomings to the models used by the research community for their analyses. Hence, this dissertation proposed an improved model and simulation framework, and the details of both are presented in Chapter 2. The failure mechanisms in STT-MRAM are then presented in Chapter 3, along with the methodology to calculate failure probabilities using the model described in Chapter 2. An optimization methodology for designing robust 1T-1MTJ STT-MRAM bit-cells based on the model and simulation framework developed in this dissertation is proposed, and is discussed in Chapter 4. Based on observations made in Chapter 4, techniques for mitigating write failure are proposed and presented in Chapter 5. Then, Chapter 6 discusses alternative storage device structures for improving STT-MRAMs. Design issues with the conventional STT-MRAM storage device are also discussed, and an alternate structure for the storage device is proposed and analyzed. Two system-level

applications that are enabled by exploiting the characteristics of STT-MRAM are then discussed in Chapter 7. Finally, Chapter 8 concludes this dissertation.

2. MODELING AND SIMULATION OF SPIN-TRANSFER TORQUE MRAM BIT-CELLS

This chapter describes the modeling of and simulation framework for magnetic tunnel junctions (MTJs) and 1T-1MTJ STT-MRAM bit-cells used in the evaluation of STT-MRAM design techniques proposed in this dissertation. Electron transport in the MTJ is modeled using the Non-Equilibrium Green's Function (NEGF) approach [48]. Magnetization dynamics in magnetic layers of the MTJ may be modeled using the Landau-Lifshitz-Gilbert (LLG) equation [49]. Magnetic field-like effects are directly captured in the LLG equation. Spin-transfer torque (STT) effects are captured in the LLG equation by adding a spin-transfer torque term. The aforementioned components of the simulation framework presented in this chapter were proposed previously and their details are presented in Appendix A (NEGF), Appendix B (LLG) and Appendix C (STT). The simulation model proposed in this dissertation allows for two different approaches to calculate the spin-transfer torque, which are also presented in the appendix. This chapter explains how the NEGF and LLG solvers are put together to simulate an entire STT-MRAM bit-cell.

2.1 Devices-to-Systems Simulation of STT-MRAM Bit-cells

Fig. 2.1 shows the role of the simulation framework proposed as part of this dissertation in the design of STT-MRAM bit-cells. Prior to device fabrication, known material parameters at the device level may be used to determine the MTJ I-V characteristics via the NEGF method discussed in Appendix A. Since spin-transfer torque characteristics may not be known prior to device fabrication, the MTJ spin-transfer torque characteristics may also be calculated using the NEGF method [50, 51] as explained earlier. The bit-cell may then be simulated to evaluate its performance before

the device is fabricated. The proposed simulation framework uses I-V characteristics for access transistors (ATx) together with the MTJ characteristics for transient simulation of the bit-cell. The advantage of the approach proposed in this dissertation is that device fabrication is not required for obtaining an initial estimate of bit-cell performance. Once a device is selected for fabrication, its characteristics calculated using the NEGF method may be verified with experimentally measured data to calibrate the simulator and obtain an accurate evaluation of its performance using the rest of the simulation framework. For example, the simulation results presented in this dissertation were obtained by calibrating both the NEGF based electron transport simulator and LLGS magnetization dynamics simulator with experimentally reported results as shown in Section 2.2.2.

In order to perform a transient simulation of an STT-MRAM bit-cell, the NEGF equations for electron transport, LLG equations for magnetization dynamics, and Kirchoff's circuit equations need to be solved simultaneously. Furthermore, the coupled equations are highly non-linear and it is often difficult to obtain analytical solutions to the equations. Hence, numerical methods are used to simultaneously solve all

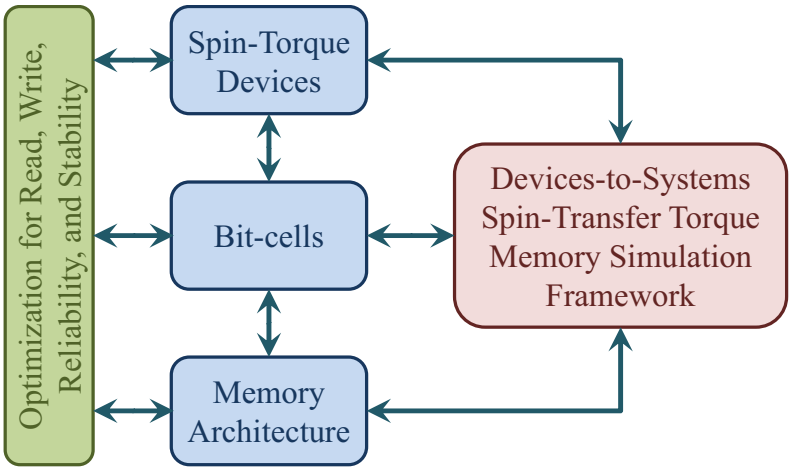


Fig. 2.1. Illustration of the role our proposed simulation framework in the STT-MRAM design and optimization process.

the equations. The simulation of STT-MRAM bit-cells in the proposed simulation framework will now be described using an example.

2.2 Simulation of the 1T-1MTJ STT-MRAM Bit-cell

The circuit model proposed for the STT-MRAM bit-cell in this dissertation is shown in Fig. 2.2. The bit and source line drivers are modeled as ideal voltage sources with output resistances R_{BS} and R_{SS}, respectively. With this circuit model, the strength of the bit and source line drivers can be controlled by varying R_{BS} and R_{SS}. Small output resistances ($\sim 1\ \Omega$) are used for strong drivers and large output resistances ($\sim 10\ M\Omega$) are used to put the driver in the high impedance state (or high-Z). This is useful for analyzing the use of voltage and current sensing schemes for reading 1T-1MTJ STT-MRAM bit-cells. The word line driver is modeled as an ideal voltage source. Stray capacitances on the bit and source lines, and the internal node, are included as C_{BL}, C_{SL} and C_{INT}, respectively. As reported in [21, 36] the electrical behavior of an MTJ is like that of a variable resistor and is modeled as R_{MTJ} in this dissertation. The circuit equations for the bit-cell model are

$$\frac{dV_{BL}}{dt} = \frac{1}{C_{BL}}\left(G_{MTJ}V_{INT} + \frac{V_{BD}}{R_{BS}} - \left(\frac{1}{R_{BS}} + G_{MTJ}\right)V_{BL}\right) \tag{2.1}$$

$$\frac{dV_{INT}}{dt} = \frac{1}{C_{INT}}\left(G_{MTJ}\left(V_B - V_{INT}\right) - I_{MOS}\right) \tag{2.2}$$

$$\frac{dV_{SL}}{dt} = \frac{1}{C_{SL}}\left(\frac{V_{SD} - V_{SL}}{R_{SS}} + I_{MOS}\right) \tag{2.3}$$

When $R_{BS} = 0$, Eq. 2.1 is ignored and $V_{BL} = V_{BD}$. Similarly, Eq. 2.3 is ignored when $R_{SS} = 0$ and $V_{SL} = V_{SD}$.

The MTJ conductance ($G_{MTJ} = R_{MTJ}^{-1}$) needs to be modeled to solve Eqs. 2.1—2.3. Since G_{MTJ} depends on the free layer (FL) magnetization as discussed in Chapter 1, the FL dynamics needs to be solved. State-of-the-art MTJs have free layers with approximate dimensions of 50 nm × 50 nm × 3 nm [36] and it has been shown that the macro-spin approximation adequately captures their magnetization dynamics [36]. Thus, the FL of the MTJ may be modeled as a mono-domain magnet.

In this dissertation, the FL of the MTJ is modeled as a macro-spin and hence, the magnetic field-term due to exchange energy may be ignored ($\nabla \widehat{m} = 0$). For an STT-MRAM bit-cell in isolation, the magnetic field-like terms that need to be accounted for are the easy axis anisotropy, the easy plane anisotropy, external applied magnetic fields, and the thermal fluctuation field. The basis vectors for the coordinate system used are $\widehat{e}_x$, $\widehat{e}_y$, and $\widehat{e}_z$. The easy axis of the magnet is along the z-axis and hence, $\widehat{m}$ is either $+\widehat{e}_z$ or $-\widehat{e}_x$ in equilibrium.

The activation energy of the magnet (E_A), which determines the thermal stability of the magnet and the retention time of the MTJ, is used to calculate the uniaxial anisotropy fields using

$$E_A = Ku_2 V \tag{2.4}$$

However, the magnet may be engineered with *in-plane anisotropy* (IMA) or *perpendicular anisotropy* (PMA). The difference between them is that the easy axes of thin film magnets with IMA lie in the plane of the thin film magnet whereas those with PMA have their easy axes pointing out-of-plane. Their anisotropies are given by

$$\textbf{IMA: } \vec{H}_{uniaxial} + \vec{H}_{easy-plane} = \frac{2Ku_2}{M_S}\widehat{m}_z - 4\pi M_S \widehat{m}_y \tag{2.5}$$

$$\textbf{PMA: } \vec{H}_{uniaxial} + \vec{H}_{easy-plane} = \left(\frac{2Ku_2}{M_S} - 4\pi M_S\right)\widehat{m}_z \tag{2.6}$$

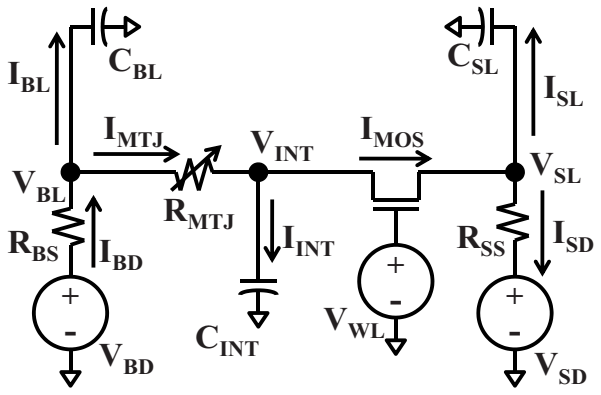


Fig. 2.2. Circuit diagram of our proposed 1T-1MTJ STT-MRAM bit-cell circuit model.

and their critical fields, H_C, are

$$H_C = \begin{cases} \frac{2Ku_2}{M_S} + 4\pi M_S \text{ for in-plane anisotropy} \\ \frac{2Ku_2}{M_S} - 4\pi M_S \text{ for perpendicular anisotropy} \end{cases} \tag{2.7}$$

These anisotropies have been experimentally observed in [12, 21, 24, 36, 52] and the origins of these anisotropies are beyond the scope of this research work.

In the STT-MRAM bit-cell, the current through the bit-cell depends on the voltages across access transistor (ATx) and R_{MTJ}, which depends on the relative angle between the pinned layer (PL) magnetization and FL magnetization and the voltage across the MTJ, V_{MTJ}. The rate at which the relative angle changes depends on the current flowing through the MTJ, I_{MTJ}. Thus, simulating the transient behavior of STT-MRAM bit-cells requires that circuit equations for the bit-cell are solved simultaneously with the equations describing the behavior of ATx and the MTJ. Fig. 2.3 shows the flow of the proposed hybrid spin-charge mixed-mode simulation framework proposed and used in this dissertation.

The I-V and C-V characteristics of ATx are given to the simulation framework either as compact models or as look-up tables generated from circuit/device simulations or from experimental data. Electrical characteristics of the MTJ are either given to the simulation framework as a compact model or calculated through NEGF simulations of electronic transport in the MTJ. Solving NEGF equations may be computationally expensive and slow [53]. Hence, the proposed simulation framework proposed allows reuse of results of NEGF simulations to speed up bit-cell simulations, as shown in Fig. 2.3. For this dissertation, R_{MTJ} calculated from NEGF simulations is encapsulated in a compact model. The key observations that allow the NEGF results to be encapsulated in a compact model are: 1) electronic transport by tunneling through a barrier has an exponential dependence on the barrier thickness, and 2) the voltage dependence of R_{MTJ} is symmetric due to symmetry in the MTJ structure. R_{AP} and R_P as a function of MTJ voltage (V_{MTJ}), the thickness of tunneling oxide (t_{OX}), and the angle between FL and PL magnetizations $\left(\theta = cos^{-1}\left(\widehat{m} \cdot \widehat{M}\right)\right)$ may

then be calculated using $R_P = R_{MTJ}(\theta = 0)$ and $R_{AP} = R_{MTJ}(\theta = \pi)$. Based on these observations, R_{AP} and R_P as functions of V_{MTJ} and t_{OX} may be individually fitted to

$$R_{MTJ} \propto \left(e^{a_0 t_{OX} + b_0} + \sum_{m=1}^{c} (-1)^{m-1} V_{MTJ}^{2m} e^{a_m t_{OX} + b_m} \right)^{-d} \tag{2.8}$$

where a_m, b_m, c and d are fitting parameters, and

$$R_{MTJ}(\theta) = \left(\frac{2R_{AP}R_P}{(R_{AP} + R_P) + (R_{AP} - R_P)\, cos\ \theta} \right) \tag{2.9}$$

Using the MTJ electrical characteristics, the spin-transfer torque in the MTJ can either be computed through Slonczewski's treatment of spin-transfer torque or through NEGF equations. The computation of spin-transfer torque is discussed in more detail

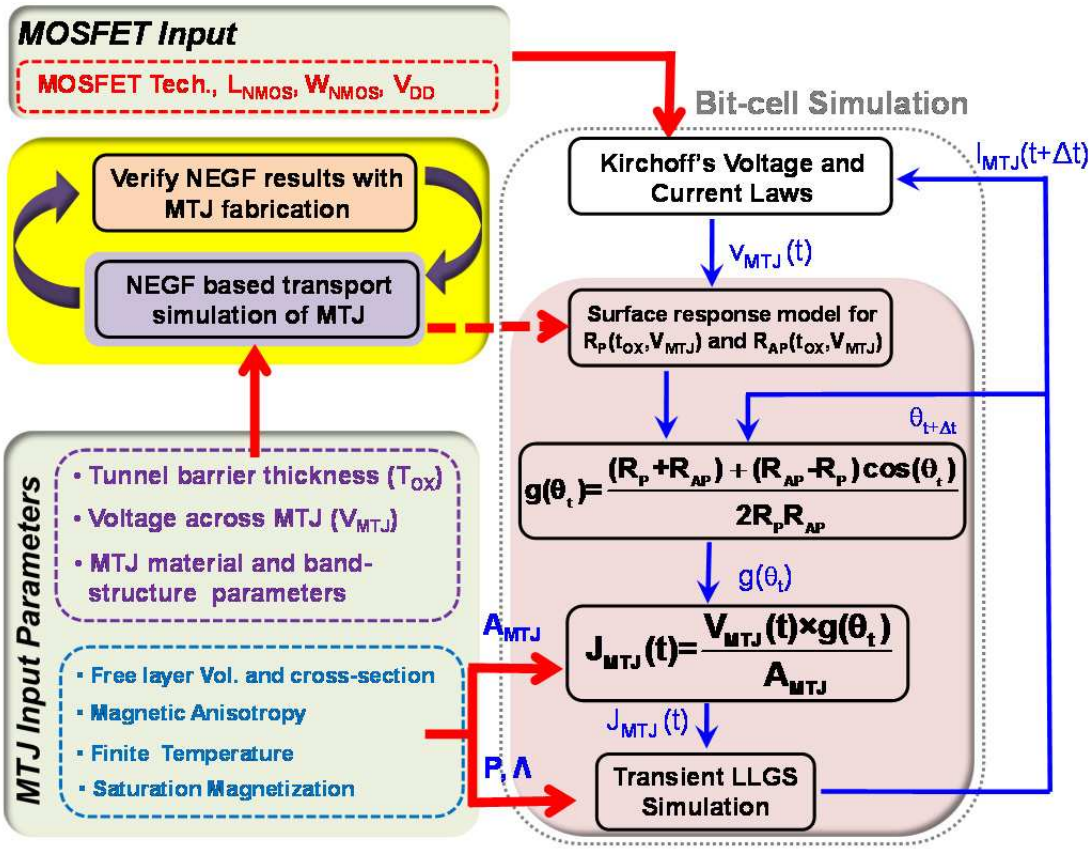


Fig. 2.3. Flow of the simulation framework proposed in this dissertation for STT-MRAM.

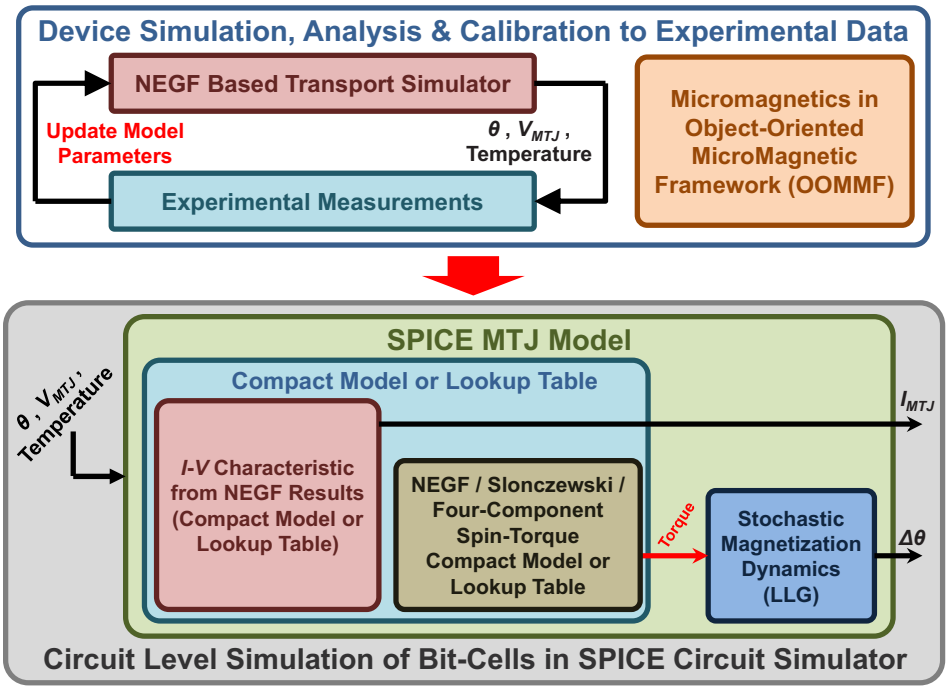


Fig. 2.4. The structure of the SPICE compatible model for the MTJ developed as part of this dissertation. The *I-V* characteristics of the MTJ is given to this model as a Verilog-A compact model. A subcircuit block for simulating the LLG equation is included in the SPICE model for the MTJ and parameters of OOMMF simulations may be given to it for SPICE simulations of magnetization dynamics.

in Appendix C. The simulation framework then numerically obtains the transient solution to the bit-cell dynamics by iteratively solving the simultaneous equations for the circuit model as well as for the magnetization dynamics.

2.2.1 Simulating magnetization dynamics in SPICE

Although the simulation framework proposed earlier in this chapter is able to perform full transient simulation of STT-MRAM bit-cells, a SPICE compatible model of the MTJ is desired so that the STT-MRAM bit-cells may be simulated directly in the SPICE circuit simulator using available SPICE models for the ATx. Hence, a SPICE compatible model for the MTJ was developed as part of this dissertation

to enable simulation of magnetization dynamics in the SPICE circuit simulator. As shown in Fig. 2.4, the I-V characteristics of the MTJ calculated using the NEGF solver maybe included as a compact model or as a look-up table. A compact model for the I-V characteristics of the MTJ was used in this disseration for simulation of circuits consisting of an MTJ. A subcircuit block is included in this SPICE model for solving magnetization dynamics in the MTJ using Eq. B.1. Hence, parameters from OOMMF simulations may be exported to this SPICE model. The subcircuit block for solving magnetization dynamics solves Eq. B.1 in spherical coordinates instead of Cartesian coordinates. Each component of the left-hand side of Eq. B.1 is represented as a node voltage on the positive terminal of a capacitor. The negative terminal of the capacitor is connected to ground. Each term of the right-hand side of Eq. B.1 is represented as a dependent current source that drives current from ground into the positive node of the capacitor representing the corresponding vector component on the left-hand side of Eq. B.1. The SPICE compatible model of the MTJ developed and used in this dissertation is compatible with HSPICE [37] and has been made available to the public on the NanoHub.org web site [54].

2.2.2 Model calibration, benchmarking and simulation results

The micromagnetic simulator in the simulation framework proposed in this dissertation was benchmarked against a gold standard micromagnetic simulator called the Object-Oriented MicroMagnetic Framework or OOMMF [55]. OOMMF simulates only micromagnetics and as such, is not suitable for simulating STT-MRAM bit-cells in which transient simulation of access transistors is required. Fig. 2.5 compares the single spin simulation results of the proposed simulator with the results returned by OOMMF. The simulations used the following parameters for the mono-domain magnet: $M_S = 850$ emu/cm^3, $\alpha = 0.03$, $\gamma = 17.6$ MHz/Oe, $T = 300$ K, $E_A = 40k_BT$, $t_{FL} = 2.1$ nm, 100 nm $\times$ 100 nm cross-sectional area, $P_L = P_R = 0.4$ and $\Lambda_L = \Lambda_R = 2$. The current flowing through the magnet is 400 μA for AP to P

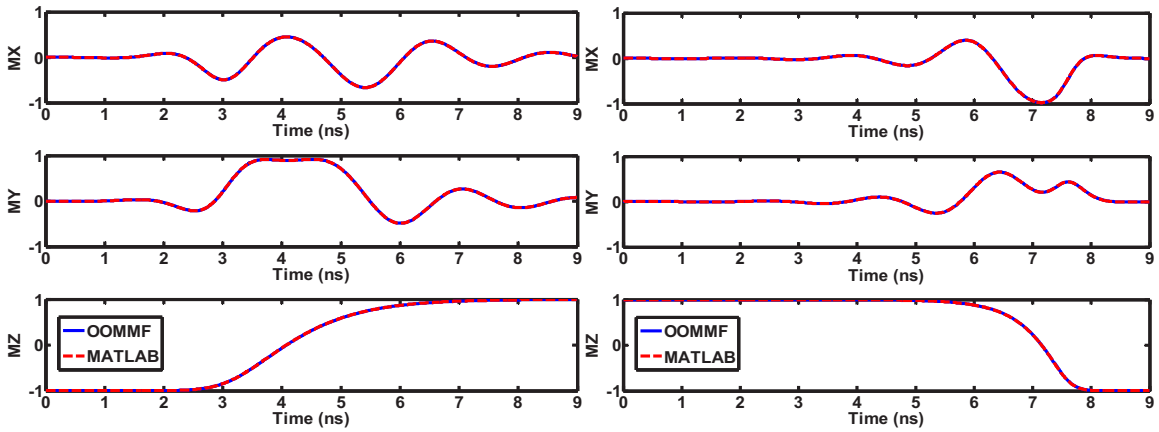


Fig. 2.5. Magnetization dynamics simulation results for a magnet driven by a constant spin-transfer torque current in OOMMF and in our simulation framework

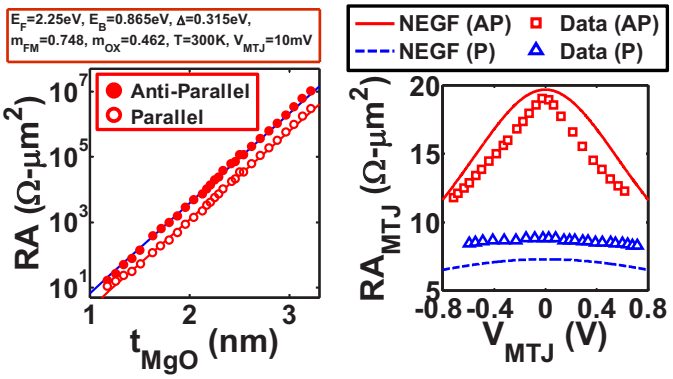


Fig. 2.6. (Left) Plot of resistance-area (RA or RA_{MTJ}) product of MTJ versus oxide thickness as obtained by our simulation framework (lines) and as reported in [13] (dots and circles). Parameters of our simulation are shown inset. (Right) The RA_{MTJ} vs. applied voltage (V_{MTJ}) at $t_{MgO} = 1.15$ nm.

switching and 800 μA for P to AP switching. Fig. 2.5 shows that the calculated magnetization trajectory of the proposed simulator completely matches that calculated by OOMMF.

Next, the NEGF solver was benchmarked to experimentally reported results in [13]. The results obtained were in reasonable agreement with the reported mea-

Table 2.1.
LLGS Paramters for 1T-1MTJ STT-MRAM Bit-cell Simulation

Access Transistor	$W = 150$ nm, 45 nm bulk CMOS
V_{DD}, V_{WRITE}	1.0 V, 1.0 V
Activation Energy, E_A	$56k_BT$, $T = 300$ K
γ, α	17.6 MHz/Oe, 0.028
Saturation Magnetization, M_S	700 emu/cm^3
Free Layer Dimensions	$\pi \times (25 \text{ nm})^2 \times 1.4$ nm
$\overrightarrow{STT}$ Fitting Parameters, P, Λ	$P_{PL} = 0.8$, $P_{FL} = 0.3$, $\Lambda_{PL} = \Lambda_{FL} = 2$

surements using the parameters $E_F = 2.25$ eV, $E_B = 0.865$ eV, $\Delta = 0.315$eV, $m_{OX} = 0.462m_0$, $m_{FM} = 0.748m_0$, $a_{OX} = a_{FM} = 0.3$ nm. Values of these parameters are within the expected range for the materials used in the MTJ. The results of resistance-area (RA) product calculated in NEGF versus oxide thickness (t_{MgO}) are shown together with experimentally reported results in Fig. 2.6. V_{MTJ} is 10 mV at temperature $T = 20$ K. Results for MTJ in P and AP configurations are plotted separately. The dependence of the RA product (RA_{MTJ}) on V_{MTJ} at $t_{MgO} = 1.15$ nm are also graphed in Fig. 2.6.

Now that the NEGF solver and LLG solver of the proposed simulator are successfully benchmarked, simulation of 1T-1MTJ STT-MRAM bit-cells is done to validate the simulation framework. MTJ I-V characteristics at $T = 300$ K were generated using the same parameters shown in Fig. 2.6. The bit-cells simulated here are similar to those reported in [36]. Since MTJ torque characteristics were not reported in [36], spin-transfer torque in the MTJ was modeled using the Slonczewski approach described in Appendix C. Table 2.1 shows the parameters used for simulating the bit-cells. Since the bit-cell configuration was not published, bit-cells with "standard connection" (SC) and bit-cells with "reversed connections" (RC) were simulated. Details of SC and RC bit-cells were discussed in Chapter 1. Fig. 2.7 shows the transient bit-cell currents and MTJ current densities during bit-cell switching. The switching

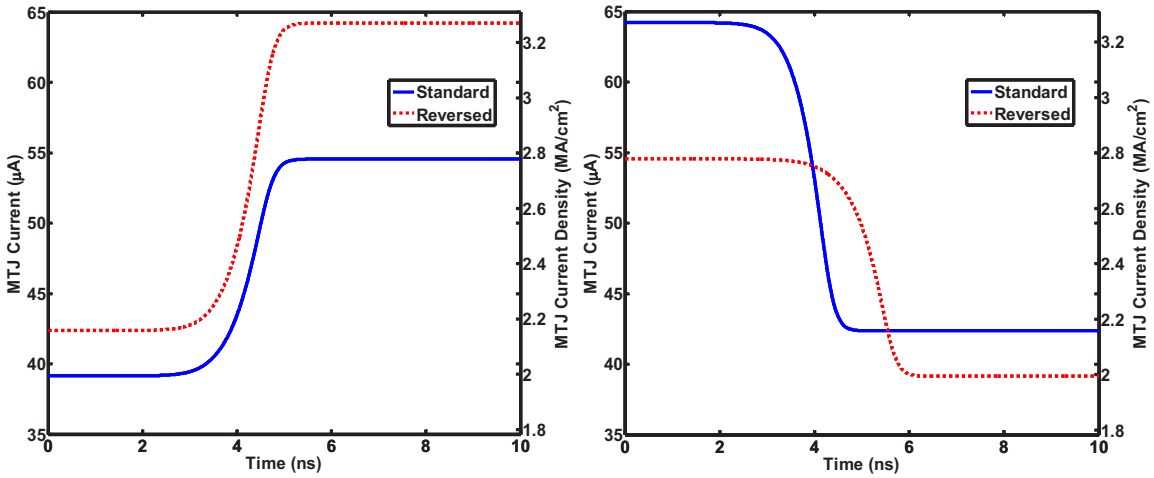


Fig. 2.7. Graphs of MTJ current and current density (bit-cell current) during bit-cell switching. (left) AP to P switching and (right) P to AP switching for SC and RC bit-cells.

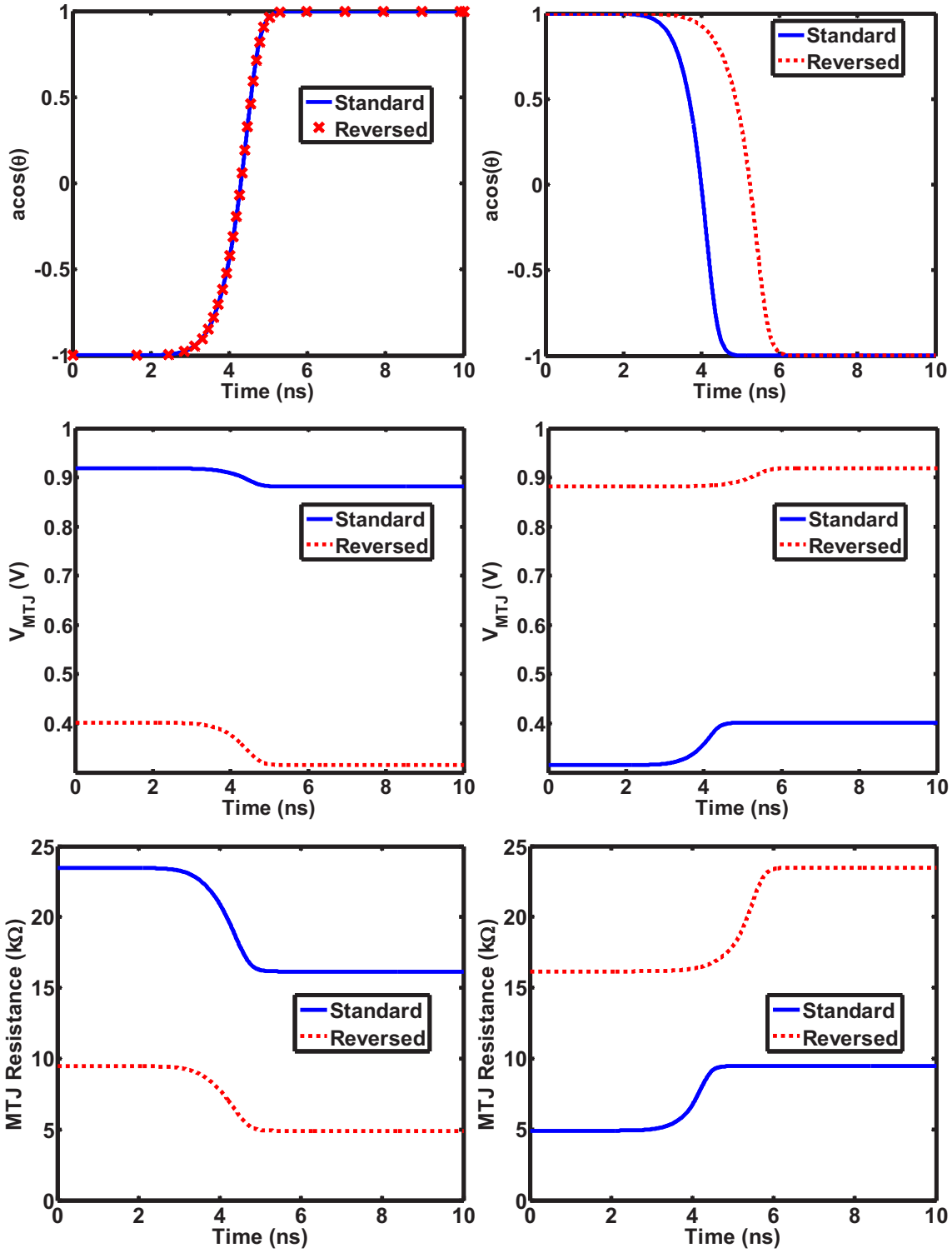


Fig. 2.8. Graphs of MTJ configuration, voltage and resistance during bit-cell switching. (left) AP to P switching and (right) P to AP switching for SC and RC bit-cells.

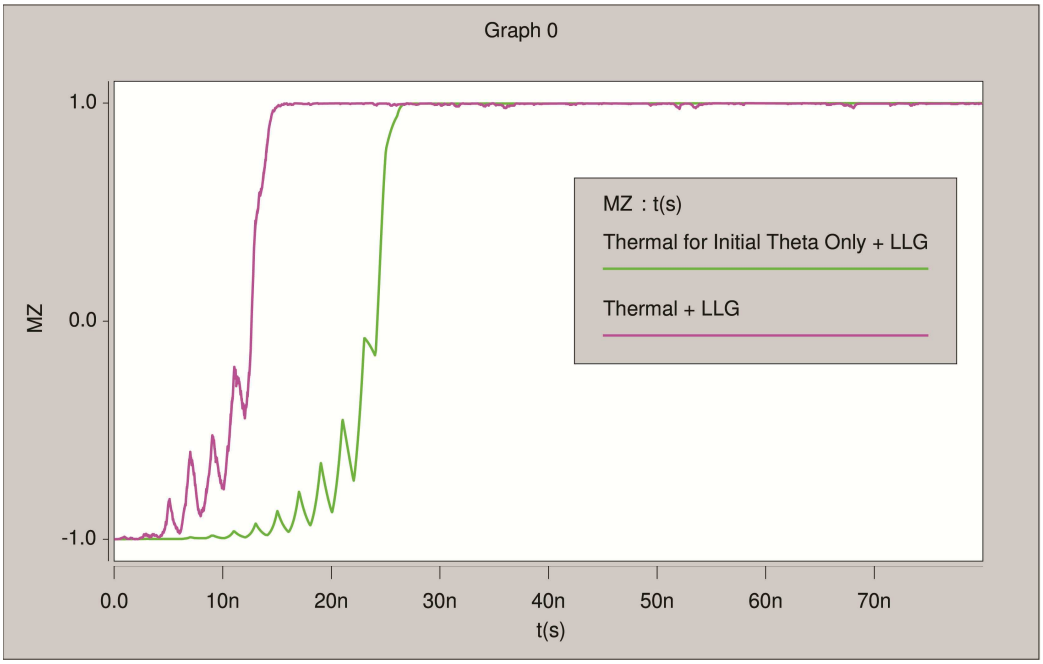


Fig. 2.9. Transient simulation of consecutive fast read operations (1 ns, V_{READ} = 1.0 V) in SPICE to compare the effect of including thermal fluctuation field on simulation results. A complete simulation shows much earlier onset of disturb failure.

delay (t_{delay}) defined here is the time taken from the beginning of MTJ switching current flow to the time when the FL magnetization is 90° relative to the PL magnetization. Other bit-cell transients are shown in Fig. 2.8. These simulation results show worst-case bit-cell switching time of $\sim$ 4.5 ns for SC bit-cells and $\sim$ 5.5 ns for RC bit-cells, which are in good agreement with experimental results reported in [36].

Finally, using HSPICE, a bit-cell is simulated in which the voltage applied across it for read operations is V_{READ} = 1.0 V and the read pulse is applied for only 1.0 ns. Thermal fluctuations were considered in one simulation but not in the other, except for only a small initial angle which simulates effects due to non-zero temperature. The second case is similar to a simulation using the models proposed in [32] and [35]. Note that even though magnetization dynamics is not captured in the model proposed in [35], the stochastic nature of switching is captured using a decision block in the model. In the case of repeated reads at high V_{READ}, the decision block needs to

capture the correlation of switching probabilities between successive read operations. The inclusion of such correlations may be very difficult in the model proposed in [35]. As graphed in Fig. 2.9, a full bit-cell simulation which considers all effects predicts earlier onset of disturb failure than a simulation that excludes thermal fluctuation field.

2.3 Summary

In this chapter, the modeling and simulation of 1T-1MTJ STT-MRAM bit-cells is described. The simulation framework proposed in this chapter simulates MTJ electron transport using the Non-Equilibrium Green's Function formalism, MTJ magnetization using the Landau-Lifshitz-Gilbert equation for magnetization dynamics, spin-transfer torque using Slonczewski's model, and compact models for the access transistor. Equations for NEGF based electron transport, LLG dynamics, spin-transfer torque and circuit behavior are solved simultaneously during transient simulation of STT-MRAM bit-cells in the proposed simulation framework. Results of bit-cell simulation obtained after calibration of the proposed simulation framework were in good agreement with experimentally reported results.

3. IMPACT OF PROCESS VARIATIONS ON STT-MRAM

A failure model for 1T-1MTJ STT-MRAM bit-cells is proposed in this chapter. The types of failures that may occur in 1T-1MTJ STT-MRAM bit-cells are also discussed. Arguments for how each type of failure may occur will be presented, using example distributions of bit-cell currents and bit-cell current densities. After discussing the origins of the failures, a methodology for determining the failure probability of each type of failure, without assuming any distributions for bit-cell currents and current densities, is proposed. The proposed methodology places assumptions only on the variations in the oxide thickness and cross-sectional area of the MTJ and transistor variations captured in SPICE models for the access transistor.

3.1 Types of Failures in 1T-1MTJ STT-MRAM Bit-cells

Recall that as described in Chapter 1, the MTJ has two configurations: the parallel (P) and anti-parallel (AP) configurations corresponding to low ($R_P = R_L$) and high ($R_{AP} =_H$) MTJ resistance, respectively. Variations in MTJ tunnel oxide thickness, t_{MgO}, and MTJ cross-sectional area, A_{MTJ}, due to process variations affect the MTJ resistance, R_{MTJ}. This results in statistical distributions for R_L and R_H. Variations in R_{MTJ} affect the ability to write into the bit-cell, the ability to correctly sense R_{MTJ} of the bit-cell, and the ability of the bit-cell to retain its state when it is being read. *Write failures* occur when the MTJ in the bit-cell cannot be switched between AP and P configurations. This may occur due to the access transistor (ATx) having a higher threshold voltage (V_T), t_{MgO} being too thick, or other factors that cause the current density through the MTJ to fall below the critical switching current density, J_C, during write operations. Failures during read operations occur when R_{MTJ} is incorrectly determined (*decision* failure) or when the MTJ configuration is acci-

dentally switched (*disturb* failure). A model for determining read and write failures was proposed in [38] and this work extends the same model for determining bit-cell failures of the "standard-connected" (SC) and "reverse-connected" (RC) 1T-1MTJ STT-MRAM bit-cell configurations. In the analysis performed, $N = 10^4$ transistor I_D–V_{DS} characteristics were obtained by Monte Carlo simulations in HSPICE and used as the characteristics of bit-cell access transistors. N may be increased to improve accuracy of results but results changed by less than 5% when N is increased to 5×10^4 in this work. Hence, $N = 10^4$ was used to speed up the analysis. MTJ conditions for read and write failures were then separately determined to calculate the respective failure probabilities of the bit-cell.

3.1.1 Write failure

Write failure occur in bit-cells that have write current densities lower than J_C because MTJ resistance (R_{MTJ}) is too large for ATx. This may occur because the ATx width is too small, the V_T of ATx is too large, t_{MgO} is too large, A_{MTJ} is too small, or a combination of factors. Under process variations, the distribution of write current density through the MTJs may look like the Gaussian distributions illustrated in Fig. 3.1(a). However, the distribution of bit-cell write current densities need not be Gaussian and will not affect the optimization methodology proposed in this chapter. The write current density in some MTJs of a particular A_{MTJ} may fall below J_C [see the vertical lines in Fig. 3.1(a)] required for switching those MTJs within the target write delay. The probability that the MTJ is unable to switch within the write delay is the *write failure probability*, P_{WRITE}. For each A_{MTJ}, J_C is determined and using N transistor I_D–V_{DS} (obtained using Monte Carlo simulations in HSPICE), the voltage across the MTJ (V_{MTJ}) is determined from the D.C. load line analysis shown in Fig. 3.1(b). Next, the maximum R_{MTJ} (and the corresponding maximum t_{MgO}, $t_{MgO-MAX}$) that allows the MTJ to be successfully written is calculated. Hence, any bit-cell having an MTJ with the same A_{MTJ} but a thicker t_{MgO} will not be success-

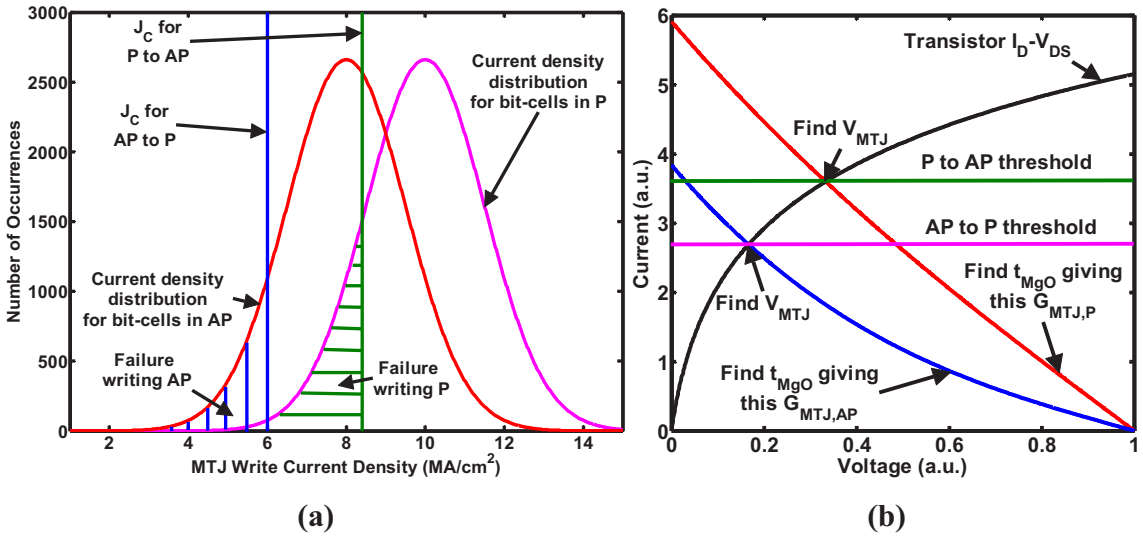


Fig. 3.1. (a) Illustration of current densities through MTJs of 1T-1MTJ STT-MRAM bit-cells during write operation under process variations. The distribution on the left represents bit-cells switching from AP to P and the one on the right for bit-cells switching from P to AP. Some bit-cells may have current densities less than J_C and thus, will not complete switching in the required write time. (b) D.C. load line used to calculate the maximum t_{MgO} that allow successful write using a particular transistor.

fully written within the target write delay. Note that because of the bi-directional write current requirement, the calculation of $t_{MgO-MAX}$ needs to be done for write '0' operations (denoted as $t_{MgO-MAX-0}$) and for write '1' operations (denoted as $t_{MgO-MAX-1}$). Thus, P_{WRITE} for the bit-cell is the probability that t_{MgO} is larger than that for the maximum R_{MTJ} calculated from the D.C. load line analysis, and may be calculated as

$$P_{WRITE} = \frac{1}{N} \sum_{\substack{all\ transistor \\ I-V}} P\left(t_{MgO} \geq min\left(t_{MgO-MAX-0}, t_{MgO-MAX-1}\right)\right) \qquad (3.1)$$

The simulations and calculations were repeated for SC and RC bit-cell configurations (which were presented in Chapter 1) to obtain their respective P_{WRITE}.

3.1.2 Read-disturb failure

Under process variations, the read current densities of 1T-1MTJ STT-MRAM bit-cells may have the Gaussian distributions as illustrated in Fig. 3.2(a). The illustrated bit-cells have read currents flowing in the parallelizing direction. The read current density of bit-cells with MTJs that have very low RA product (due to thinner t_{MgO} or other reasons) can be higher than J_C. Hence, the MTJ may be unintentionally written into during read operations. The probability that the MTJ in a bit-cell is unintentionally written into during read operations is $P_{READ-DISTURB}$ and is calculated in the same manner as P_{WRITE}. The differences are that a successful write operation during read operation is a disturb failure and flips only one MTJ state (either P or AP). The direction of current flowing through the MTJ during read operations may parallelize or anti-parallelize the MTJ, depending on the bit-cell configuration and bit line and source line voltages. For anti-parallelizing read operations, only MTJs in P may be unintentionally flipped to AP. For parallelizing reads, only MTJs in AP may be unintentionally flipped to P. The current densities through the MTJs during read may not have a Gaussian distribution and was assumed so only for illustration. The proposed failure calculation methodology does not assume the distribution of bit-cell currents during read operations.

For a particular A_{MTJ}, the minimum R_{MTJ} that results in a read-disturb failure is first calculated. This condition is only met for a specific t_{MgO}, denoted as $t_{MgO-MIN}$. If t_{MgO} is thinner, R_{MTJ} becomes smaller and the MTJ will be written into during read. Thus, the probability t_{MgO} is thinner than $t_{MgO-MIN}$ is $P_{READ-DISTURB}$. For each A_{MTJ}, J_C is determined and using N transistor I_D–V_{DS} obtained using Monte Carlo simulations in HSPICE, V_{MTJ} is determined using the D.C. load line analysis shown in Fig. 3.2(b). Next, the maximum R_{MTJ} (and the corresponding t_{MgO}, $t_{MgO-MIN}$) that suffers read disturb when paired with each ATx calculated. Hence, any bit-cell having the same ATx and an MTJ with the same A_{MTJ} but a thinner t_{MgO} will be disturbed during read when the MTJ is in AP. Thus, $P_{READ-DISTURB}$

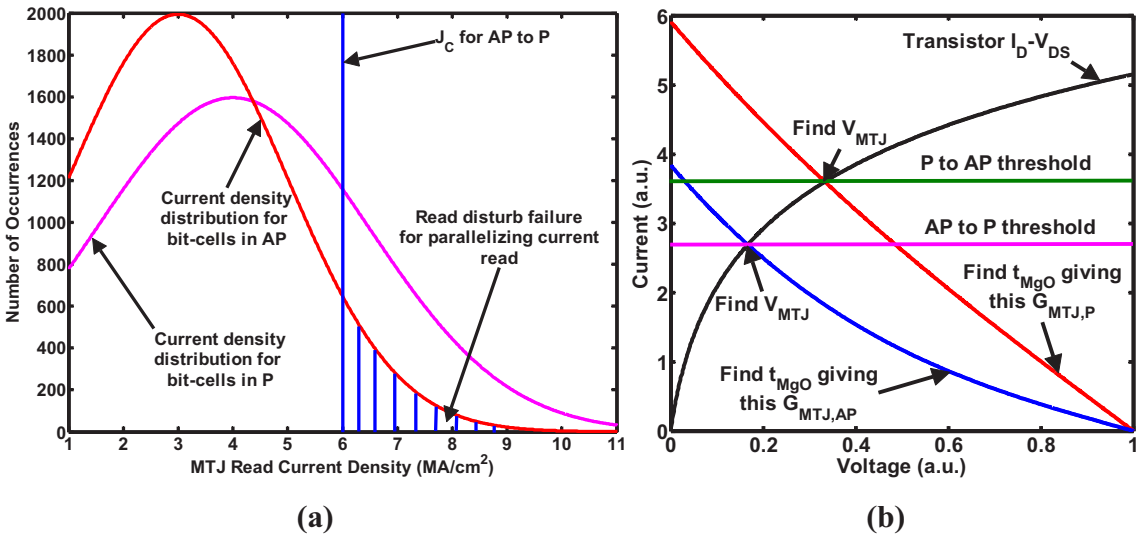


Fig. 3.2. (a) Illustration of current densities through MTJs of 1T-1MTJ STT-MRAM bit-cells during read operation under process variations. The distribution on the left represents bit-cells in AP the one on the right for bit-cells in P. When the read current is in parallelizing direction, some bit-cells may have current densities more than J_C and thus, will get switched during. (b) D.C. load line used to calculate the minimum t_{MgO} that suffers read disturb using a particular transistor.

for the bit-cell is the probability that t_{MgO} is larger than that for the maximum R_{MTJ} calculated using D.C. load line analysis, and

$$P_{READ-DISTURB} = \frac{1}{N} \sum_{\substack{alltransistor \\ I-V}} P\left(t_{MgO} \leq t_{MgO-MIN}\right) \tag{3.2}$$

Depending on the bit-cell configuration, the read current direction (and thus the bit and source line voltages) needs to be carefully chosen to minimize $P_{READ-DISTURB}$.

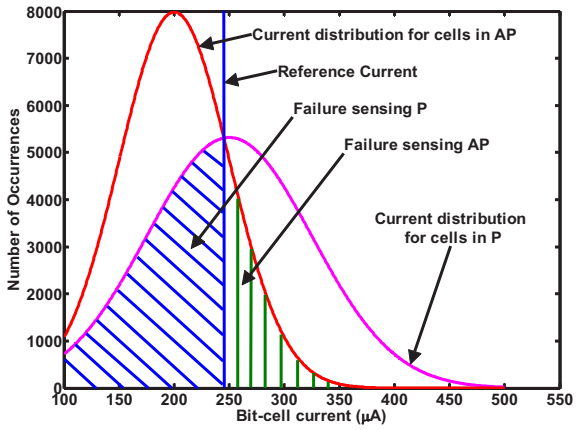
3.1.3 Read-decision failure

During STT-MRAM read operations, the voltages of the bit, source, and word lines are fixed and current flows through the bit-cell via the MTJ and into a current-sense amplifier. The sense amplifier compares the bit-cell current to a reference current (I_{REF}) to determine R_{MTJ} and hence the magnetic configuration of the MTJ.

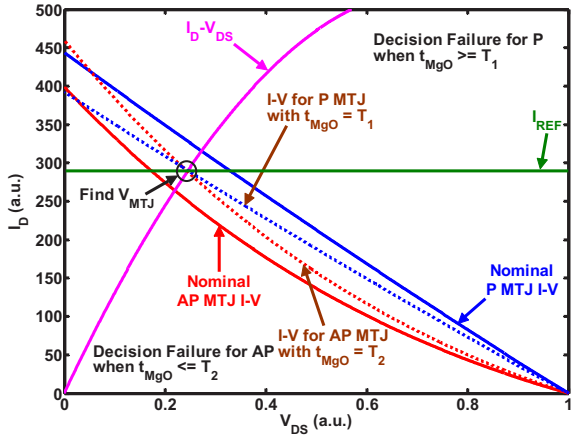
If the bit-cell current is less than I_{REF}, then $R_{MTJ} = R_H$ and the sense amplifier outputs H or '1'. If the bit-cell current is more than I_{REF}, then $R_{MTJ} = R_L$ and the sense amplifier outputs L or '0'.

Even for fixed voltages on the word, source and bit lines of the STT-MRAM bit-cell, the current flowing through the bit-cell during read operations may vary due to process variations. Fig. 3.3(a) illustrates the distribution of read currents that STT-MRAM bit-cells may have. The Gaussian distribution on the left represents the current through bit-cells when the MTJ is in AP and the distribution on the right are for same bit-cells when the MTJ is in P. Some bit-cells in P have currents less than I_{REF} and some bit-cells in AP have currents more than I_{REF}. The sense amp will not be able to correctly determine R_{MTJ} in these bit-cells. The distribution of bit-cell current does not need to be Gaussian and was assumed so only for illustration purposes. The failure calculation methodology proposed in this chapter does not assume the distribution of bit-cell currents during read.

Decision failures occur when the sense amplifier outputs H for a bit-cell in P configuration (R_L) and when the sense amplifier outputs L for a bit-cell in AP configuration (R_H). The probability that a correctly functioning sense amplifier incorrectly senses R_{MTJ} in the bit-cell is the read-decision failure, $P_{READ-DECISION}$. I_{REF} needs to be chosen to minimize $P_{READ-DECISION}$. For a bit-cell with an MTJ of a particular A_{MTJ} and configuration, a particular t_{MgO} ($t_{MgO-AP-REF}$ for MTJ in AP and $t_{MgO-P-REF}$ for MTJ in P) will result in the bit-cell current to be I_{REF}. If the MTJ is in AP, a thinner t_{MgO} will result in a smaller R_{MTJ} and a bit-cell read current higher than I_{REF}. The sense amp will incorrectly determine R_{MTJ} to be R_L during read. Thus, $P_{READ-DECISION}$ of the bit-cell is the probability t_{MgO} is lower than $t_{MgO-AP-REF}$. Similarly, if the MTJ is in P, a thicker t_{MgO} will result in a larger R_{MTJ} and a bit-cell current lower than I_{REF}. The sense amp will incorrectly determine R_{MTJ} to be in R_H during read. Thus, $P_{READ-DECISION}$ of the bit-cell is the



(a)



(b)

Fig. 3.3. (a) Illustration of MTJ read current distribution in 1T-1MTJ STT-MRAM bit-cells under process variations. The distribution on the left represents bit-cells in AP the one on the right for bit-cells in P. Some bit-cells in P may have currents less than I_{REF} and some bit-cells in AP may have currents more than I_{REF}. (b) D.C. load line used to calculate the maximum t_{MgO} that allow successful write using a particular transistor.

probability t_{MgO} is larger than $t_{MgO-AP-REF}$. Hence, $P_{READ-DECISION}$ for a specific I_{REF} may be calculated as

$$P_{READ-DECISION} = \frac{1}{N} \sum_{\substack{all\ transistor \\ I-V}} [P\left(t_{MgO} \leq t_{MgO-AP-REF}\right) + P\left(t_{MgO} \geq t_{MgO-P-REF}\right)] \tag{3.3}$$

The optimum I_{REF} ($I_{REF-OPT}$) that minimizes $P_{READ-DECISION}$ lies between the nominal bit-cell currents of bit-cell with MTJ in AP (I_{AP}) and of bit-cell with MTJ in P (I_P). In the proposed failure calculation methodology, a linear search is done between I_{AP} and I_P to find $I_{REF-OPT}$. Calculation time maybe significantly increased if the calculation of $P_{READ-DECISION}$ requires sweeping the A_{MTJ} during the linear search. Instead, an approximation is used so there is no need to sweep AMTJ. For the bit-cell with MTJ in AP (P), we use an MTJ with A_{MTJ} that is six sigma less (more) than nominal for calculation. The t_{MgO} that results in the bit-cell current to be I_{REF} is higher (lower) when calculated this way. Thus, the calculated $P_{READ-DECISION}$ is larger than actual and provides an approximate upper bound.

3.2 Total failure probability of 1T-1MTJ STT-MRAM Bit-cells

In the failure model proposed in [38], read and write failures are assumed to be independent and hence, the total failure probability of a bit-cell is the sum of read and write failures. However, the failure model proposed here indicates that write and read failures may not be independent. Bit-cells that have MTJ with excessively large t_{MgO} may have write failure as well as decision failure. Hence, the total failure probability of STT-MRAM bit-cells ($P_{FAILURE}$) may instead be calculated as

$$P_{FAILURE} = \frac{1}{N} \sum_{\substack{all\ transistor\\ I-V\ and\\ MTJ\ area}} min\left(1, P\left(t_{MgO} \geq min\left(t_{MgO-MAX}, t_{MgO-P-REF}\right)\right)\right) \times P\left(t_{MgO} \leq max\left(t_{MgO-MIN}, t_{MgO-AP-REF}\right)\right)\right) \quad (3.4)$$

3.3 Summary

In this chapter, a failure model for 1T-1MTJ STT-MRAM bit-cells is proposed. A discussion of write, read-disturb and read-decision failures and their occurrence under process variations was presented. A methodology for calculating each failure probability using D.C. load line analysis together with HSPICE Monte Carlo sim-

ulation was also proposed. The total failure probability for 1T-1MTJ STT-MRAM bit-cells calculated using the proposed methodology does not assume independence of read and write failures. Furthermore, the proposed methodology does not assume any distribution for bit-cell currents and current densities. The results of the failure analysis using the proposed methodology depend only on the distributions assumed for A_{MTJ}, t_{MgO} and I–V characteristics of ATx.

4. OPTIMIZATION OF 1T-1MTJ STT-MRAM BIT-CELLS

This chapter proposes a bit-cell optimization methodology using proper selection of bit-cell configuration and proper sizing of access transistor (ATx) to minimize failures in 1T-1MTJ STT-MRAM bit-cells. Bit-cell failure probabilities are calculated using the failure model presented in Chapter 3. Analysis of the proposed optimization technique on 1T-1R STT-MRAM bit-cells designed using 45 nm bulk CMOS and 45 nm silicon-on-insulator (SOI) technologies is also presented. The ITRS roadmap shows that by 2016, transistor gate lengths and MTJ lateral dimensions are expected to reach 16 nm and 32 nm, respectively [23]. Scaled MTJs can be engineered to meet performance requirements at iso-stability as explained in [12]. Thus, the optimization technique proposed in this chapter is used to estimate the expected iso-stability failure probabilities of 1T-1R STT-MRAM bit-cells in 2016. A 16 nm Predictive Technology Model (PTM) is used to model 16nm gate length CMOS ATx in HSPICE. The MTJ model proposed in Chapter 2 was used to model MTJs with 32 nm $\times$ 32 nm cross-section. ATx variations were simulated using variations in V_T (μ = 480mV, σ = 30mV). This chapter is organized as follows. The MTJ characteristics and assumptions in MTJ variations used in the analysis of the optimization methodology are presented first. Results of analysis performed on bit-cells simulated in 45 nm bulk CMOS, 45 nm silicon-on-insulator (SOI) and 16 nm predictive (PTM) technologies are presented next. Specifically, the impact of bit-cell read voltage (V_{READ}) on read-disturb and read-decision failures are discussed first, and it is then shown that proper selection of V_{READ} allows control over whether disturb or decision failure is the dominant form of read failure. The impact of NFET sizing on write and read failures are discussed and compared next. The heuristic for determining the optimum bit-cell configuration and NFET size is also presented.

4.1 Proposed Technique for Optimizing 1T-1MTJ STT-MRAM Bit-cells

The flow of our proposed optimization methodology is shown in Fig. 4.1. Our simulation framework is calibrated first and then used to generate MTJ resistance and switching characteristics for use in rest of the analysis. Initial NFET sizing for meeting J_C in bit-cells of all configurations (standard connection or SC, and reverse connection or RC) is done without considering process variations to determine an initial starting point for NFET sizing. I–V characteristics for $N = 10^4$ NFET (with variations) of initial NFET size are then generated using Monte Carlo simulations in SPICE. Failure probabilities for all bit-cell configurations are then calculated. The probabilities correspond to bit-cells having initial NFET sizing. NFET width swept to obtain the failure probabilities versus NFET width. The optimum NFET size for

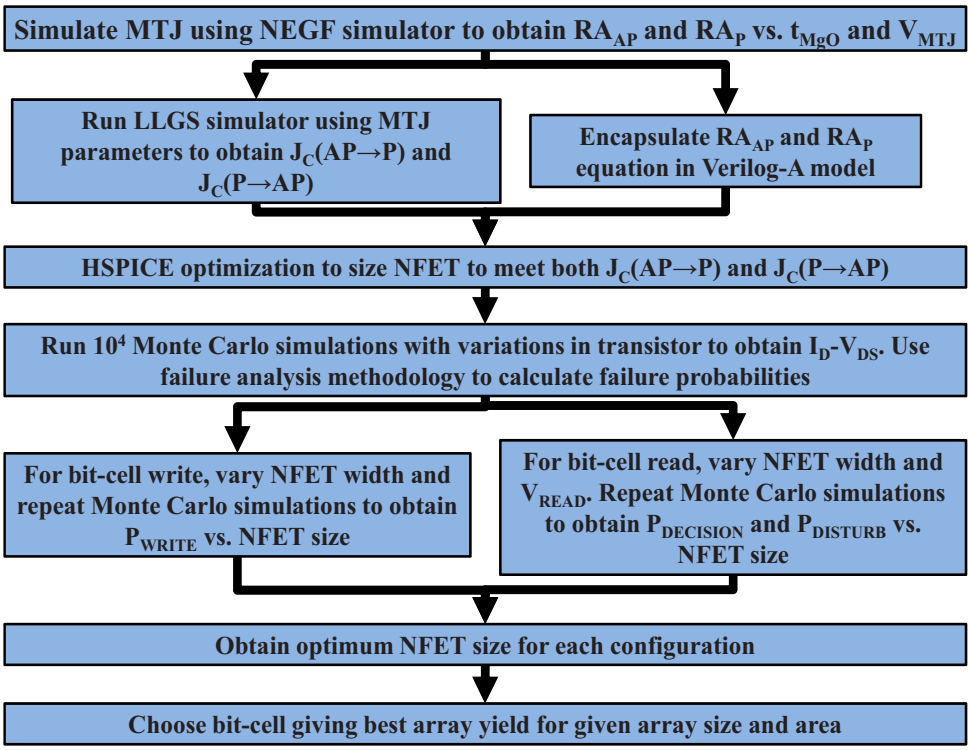


Fig. 4.1. Illustration of the flow of our proposed optimization technique.

each bit-cell configuration is determined and the bit-cell configuration that gives the best array yield for a given array size and area is selected.

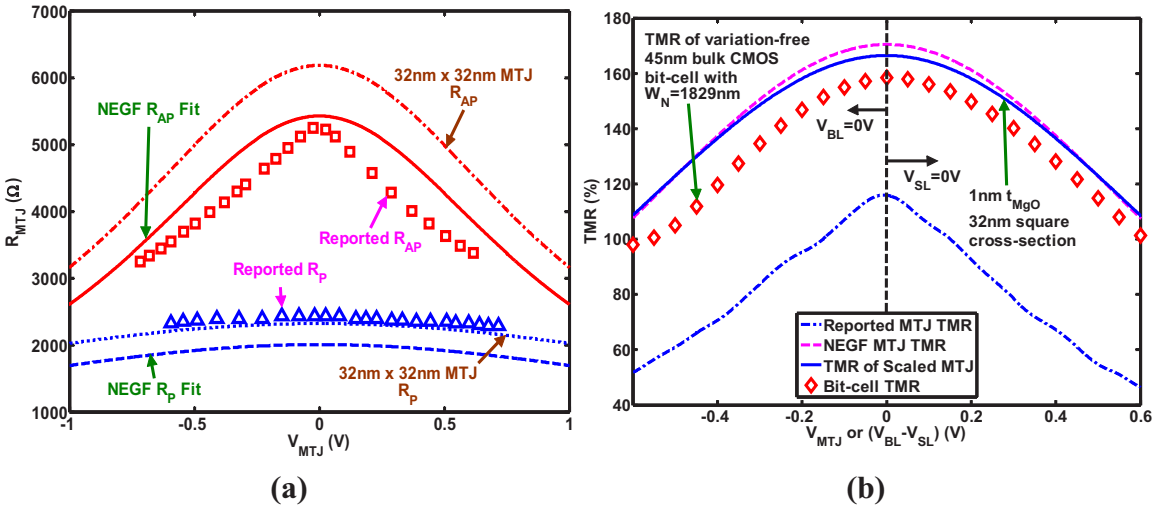


Fig. 4.2. (a) Comparisons of R_{MTJ} vs. V_{MTJ} reported in experiment [21] (squares and triangles) and from our calibrated simulation framework. (b) TMR vs. V_{MTJ} of corresponding MTJs (a). The MTJ with 32 nm × 32 nm cross-section is the scaled MTJ.

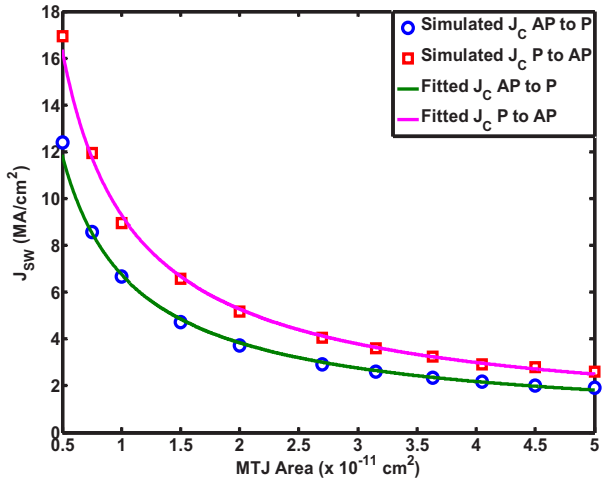


Fig. 4.3. J_{SW} (or J_C) vs. MTJ cross-sectional area of the MTJ used in our analysis.

Table 4.1.
Parameters for Simulated STT-MRAM Bit-cells

Nominal J_C(AP→P)	$\sim$ 2.35 MA/cm^2
Nominal J_C(P→AP)	$\sim$ 3.22 MA/cm^2
Nominal Free Layer Volume (Ellipse)	40 nm × 116 nm × 1.5 nm
PMA Anisotropy Energy Barrier	$E_A = 51k_BT$
Saturation Magnetization (M_S)	850 emu/cm^2
Damping Factor (α)	0.028
Gyromagnetic Factor (γ)	17.6 GHz/Oe
45 nm t_{MgO}, 16 nm t_{MgO}, t_{delay}	1.15 nm, 1.0 nm, $\sim$40 ns
V_{DD}, V_{READ}	1.0 V, 0.1 V
$V_{WRITE} = \|V_{BL} - V_{SL}\|$	1.0 V

4.2 Characteristics of MTJ Under Analysis

Our simulation framework was first calibrated to experimentally reported data and then used to generate MTJ characteristics for use in our analysis. The calibration of the proposed simulation framework using material parameters was presented earlier in Chapter 2. Fig. 4.2(a) shows the graph of MTJ resistance (R_{MTJ}) versus the voltage across the MTJ (V_{MTJ}). The MTJ characteristics reported in [21] are plotted together with the MTJ characteristics using the proposed simulation framework. The MTJ characteristics for identical MTJ dimensions (elliptical cross-section with 40 nm short axis and 116 nm long axis, and 1.5 nm free layer thickness) are in reasonably good agreement. The MTJ characteristic for an MTJ with 32 nm × 32 nm square cross-sectional area with identical oxide thickness (t_{MgO}) is also plotted in Fig. 4.2(a). The tunneling magnetoresistance ratio (TMR) versus V_{MTJ} calculated by our simulation framework and that reported in [21] are graphed in Fig. 4.2(b). The TMR of the MTJ used in our analysis is substantially higher than the TMR reported in [21]. However, the trend of TMR versus V_{MTJ} is in good agreement. The

difference in MTJ characteristics does not impact the correctness of the optimization methodology proposed in this chapter, but the optimum NFET width and bit-cell failure probabilities are affected.

As mentioned in Chapter 3, the critical current density (J_C) for switching MTJ configurations within a fixed period of time is needed for calculating failure probabilities for 1T-1MTJ STT-MRAM bit-cells. Other than J_C, other metrics commonly quoted in literature include the critical switching current (I_C) or the V_{MTJ} at which the MTJ is switched [20,21,36]. Note that I_C or J_C may also be defined as the required current or current density required to ensure MTJ switching occurs, independent of switching delay. Throughout this research, the switching time dependent definition of J_C and I_C ($I_C = J_C \times A_{MTJ}$, A_{MTJ} = MTJ cross-sectional area) is used. J_C also depends on factors such as the free layer magnetic anisotropy, applied magnetic fields on the free layer, and other factors as discussed in Chapter 2. Interestingly, J_C is independent of t_{MgO} in the model proposed in this dissertation. Since variations are assumed only in t_{MgO} and MTJ lateral dimensions, variations in J_C are due only to variations in MTJ cross-sectional area in the analysis to be presented next. Fig. 4.3 shows the graph of J_C versus A_{MTJ} for 40 ns switching delay. We have defined switching delay (t_{delay}) as the time taken for spin-transfer torque to rotate the free layer magnetization from 0° or 180° to 90°. The circles and squares in Fig. 4.3 are J_C as determined from the simulation framework. The lines in Fig. 4.3 are fitted to these data points and used to model the MTJ area dependence in J_C of the MTJ used in the analysis later. Together, Fig. 4.2 and Fig. 4.3 represent the characteristics of the MTJ used for evaluating the effectiveness of the optimization methodology proposed in this chapter. Parameters of the MTJ for bit-cells simulated in 45 nm bulk CMOS and 45 nm SOI technologies are summarized in Table 4.1. Parameters of the MTJ for bit-cells simulated in 16 nm PTM are the same except for A_{MTJ} and the nominal J_C. Also, the analysis was performed assuming $\frac{\sigma}{\mu} = 2\%$ in t_{MgO} and $\frac{\sigma}{\mu} = 5\%$ in MTJ cross-sectional area. μ for t_{MgO} and A_{MTJ} were kept at nominal values for the anal-

ysis. For 16 nm PTM technologies, NFET variations were simulated using variations in V_T ($\mu = 480$mV, $\sigma = 30$mV).

4.3 Simulation Results and Analysis of Proposed Optimization Technique

The results and analysis of optimized 1T-1MTJ STT-MRAM bit-cells in 45 nm bulk CMOS, 45 nm SOI and 16 nm PTM technologies are presented in this section. The selection of V_{READ} and the impact of V_{READ} on read failure of all the bit-cells are discussed first. After that, the impact of NFET width on bit-cell failures is presented. The results are then used to discuss the heuristic for determining optimality.

4.3.1 Selection of V_{READ}

The read failures of 1T-1MTJ STT-MRAM bit-cells in 45 nm bulk CMOS and 45 nm SOI technologies (standard V_T) are plotted against V_{READ} in Fig. 4.4(a) and

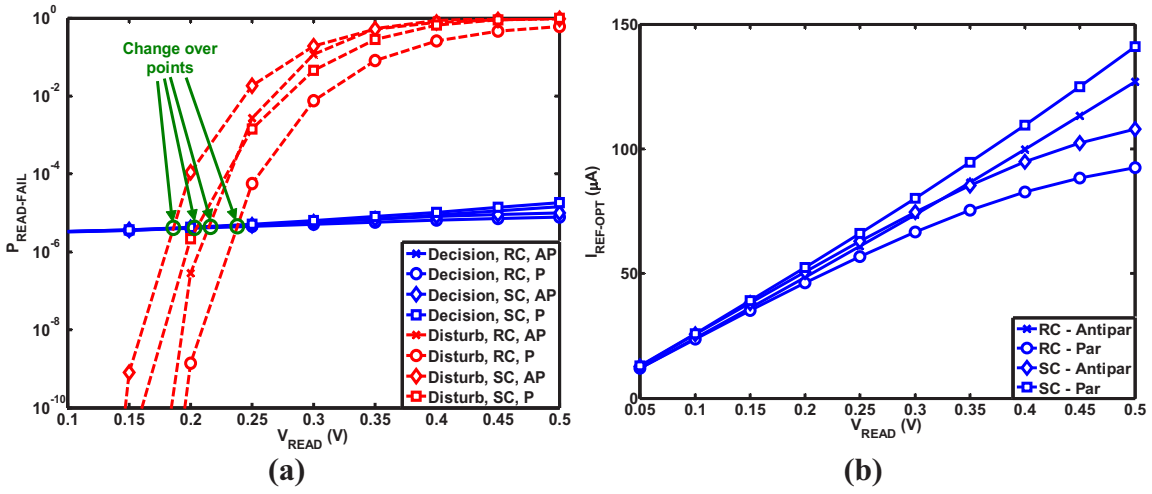


Fig. 4.4. (a) Read failures vs. V_{READ} and (b) corresponding IREF-OPT for 1T-1MTJ STT-MRAM bit-cells in 45 nm bulk CMOS technology. NFET widths are 671 nm and 405 nm for SC and RC bit-cells, respectively.

Fig. 4.5(a), respectively. NFET width is kept constant for each technology while V_{READ} is varied to determine the read failures. NFET width for SC and RC bit-cells implemented using 45 nm bulk CMOS transistors are 671 nm and 405 nm, respectively. Decision failures and disturb failures are plotted separately to show that the choice of V_{READ} determines which read failure is dominant. The change over points indicate that for the RC bit-cell, a larger V_{READ} can be used before disturb failures become dominant compared to the SC bit-cell. Also, disturb failures decrease as V_{READ} is reduced because lower V_{READ} reduces bit-cell currents and hence, lowers the current density through the MTJ during read. Decision failures decrease with lower V_{READ} because of higher TMR at lower V_{MTJ}. Degradation of TMR with V_{MTJ} (or I_{MTJ}) has been widely reported [20,21,56] and Fig. 4.2(b) illustrates the TMR degradation captured in our NEGF based MTJ model. For small V_{MTJ}, the TMR of the MTJ approaches its maximum (170%) at $t_{MgO} = 1.15$ nm. However, the bit-cell TMR is always lower than the TMR of the MTJ due to the resistance of the access transistor that appears in series with R_{MTJ}. When the transistor resistance becomes the dominant contributor to the total bit-cell resistance, distinguishability between R_P and

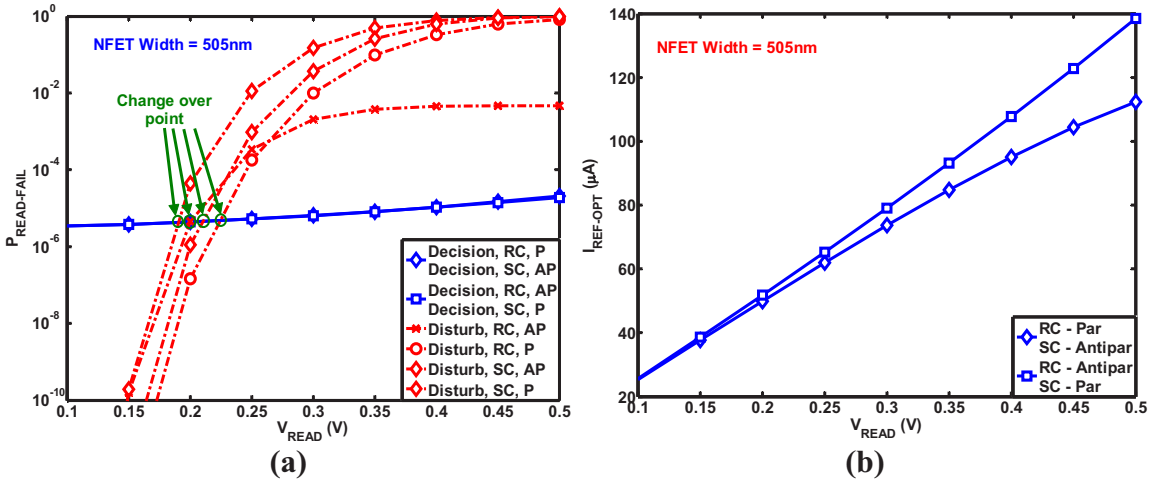


Fig. 4.5. (a) Read failures vs. V_{READ} and (b) corresponding $I_{REF-OPT}$ for 1T-1MTJ STT-MRAM bit-cells in 45 nm SOI technology.

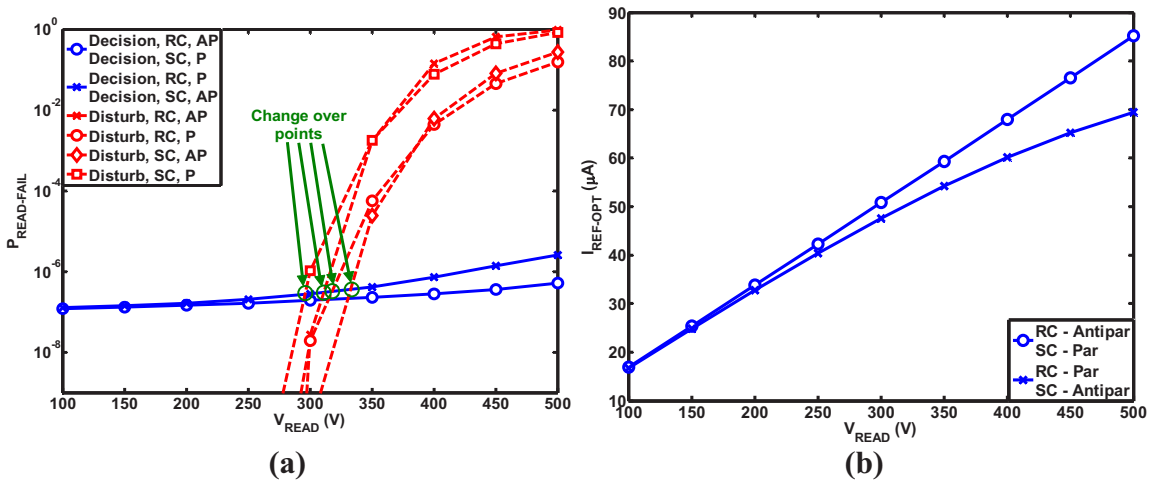


Fig. 4.6. (a) Read failures vs. V_{READ} and (b) corresponding $I_{REF-OPT}$ for 1T-1MTJ STT-MRAM bit-cells in 16 nm PTM technology.

R_{AP} is reduced. The corresponding $I_{REF-OPT}$ for the calculated $P_{READ-DECISION}$ in Fig. 4.4(a) and Fig. 4.5(a) are graphed in Fig. 4.4(b) and Fig. 4.5(b), respectively.

The graph of 16 nm PTM based STT-MRAM bit-cell read failures versus V_{READ} at fixed NFET width is shown in Fig. 4.6(a). The same trends found in read failures for STT-MRAM based on 45 nm CMOS technologies are also observed. However, V_{READ} at which disturb failures become more dominant are much higher than 45 nm CMOS based STT-MRAM bit-cells (likely due to higher V_T). Thus, decision failure is expected to remain the dominant failure in future STT-MRAM.

4.3.2 Effect of NFET sizing and proposed heuristic for optimality

The graph of P_{WRITE} versus NFET width for bit-cells in 45 nm transistor technologies is shown in Fig. 4.7. As the NFET width increases, more of V_{WRITE} is dropped across the MTJ (i.e., V_{MTJ} increases). Thus, the fundamental limit of bit-cell write failure can be calculated by ignoring the NFET and assuming $V_{MTJ} = V_{WRITE}$. This is analogous to assuming an infinitely wide NFET. For our MTJ and choice of

V_{WRITE}, the fundamental limit of bit-cell write failure is less than 10^{-20}. The NFET width needed to achieve this is very large and unfeasible for high-density STT-MRAM

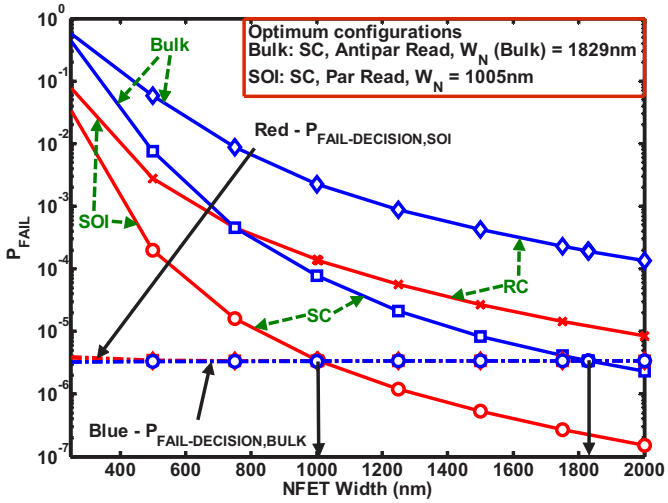


Fig. 4.7. Write and read failures vs. NFET width for bit-cells in 45 nm bulk CMOS and 45 nm SOI technologies. Optimum NFET width occurs when write and decision failure probabilities are equal. Failure probability at the optimum width is $\sim 3.4 \times 10^{-6}$.

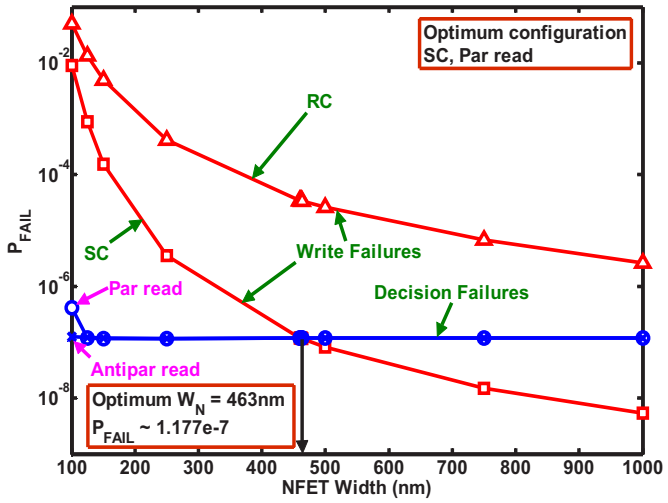


Fig. 4.8. Write and read failures vs. NFET width for bit-cells in 16 nm PTM technology. Optimum NFET width occurs when write and decision failure probabilities are equal. Failure probability at optimum width is $\sim 1.18 \times 10^{-7}$.

arrays. Furthermore, read decision failures become the dominant failure beyond a certain NFET width.

In order to determine the optimum NFET width, the dominant bit-cell failure needs to be determined. As shown in Fig. 4.4(a) and Fig. 4.5(a), decision failures are the dominant read failure at sufficiently small V_{READ}. Since the decision and disturb failures can be controlled independent of write failures by setting V_{READ}, V_{READ} is set such that read failures are as small as possible and are dominated by decision failures ($V_{READ} = 0.1$V). Write failures and decision failures are then compared with varying NFET width (Fig. 4.7). Our simulations show that write failures are higher than read failures over a wide range of NFET widths. $P_{READ-DISTURB}$ for both 45 nm bulk CMOS and 45 nm SOI technologies are less than 2×10^{-10} over the entire range of NFET width. Compared to the RC bit-cell, the NFET width at iso-P_{WRITE} for SC bit-cell is much smaller. Thus, the SC bit-cell has better yield at iso- bit-cell area as compared to the RC bit-cell. Also, write failures are lower than decision failures when the NFET is wide enough. Beyond that width, decision failure dominates and increases with increasing NFET width. The optimum NEFT width is the one at which decision failures are equal to write failures. Since array area and bit-cell density are the primary concerns for memory arrays, the optimum bit-cell configuration is the one that requires the smallest NFET width. The optimum NFET width is about 1829 nm and about 1005 nm for SC bit-cells in 45 nm bulk and 45 nm SOI CMOS technologies, respectively. The failure probabilities of the optimum bit-cells in 45 nm bulk CMOS and 45 nm SOI technologies are both $\sim 3.4 \times 10^{-6}$. The optimum read current configuration is to have anti-parallelizing read for SC bit-cells implemented in 45 nm bulk CMOS technology, whereas that for SC bit-cells implemented in 45 nm SOI is to have parallelizing read. $I_{REF-OPT}$ are 27.83 μA and 27.29 μA for bit-cells implemented in 45 nm bulk CMOS and 45 nm SOI technologies, respectively.

Note that V_{WRITE} has been set to be V_{DD} in our bit-cells since STT-MRAM bit-cells are anticipated to be embedded close to the processor core where higher I/O voltages are not readily available. A higher V_{WRITE} allows more write current density

through the MTJ during write. Hence, write failures may be reduced by increasing V_{WRITE} if higher supply voltages are available. Higher V_{WRITE} also allows the NFET width to be reduced at iso-P_{WRITE} at the expense of increased power dissipation. However, decision failure may increase slightly when NFET width is reduced, as shown in Fig. 4.7.

P_{WRITE} versus NFET width for bit-cells in 16 nm PTM technology are shown in Fig. 4.8. The trends observed are similar to those in the bit-cells in 45 nm transistor technologies. The write failure improvement diminishes with increasing NFET width. Write failure and decision failure versus NFET width are compared with $V_{READ} =$ 100 mV (decision failure is the dominant read failure under this condition). NFET width of the RC bit-cell is limited by AP to P switching (source degenerated NFET). However, compared to the RC bit-cell, NFET width at iso-P_{WRITE} for SC bit-cell is much smaller. Thus, SC bit-cells implemented in 16 nm PTM shows better yield at iso- bit-cell area. The optimum NFET width occurs at the point where decision failure just dominates write failure (463 nm). The bit-cell failure probability is $\sim 1.18 \times 10^{-7}$ and $I_{REF-OPT}$ for sensing the MTJ resistance is 23.02 μA. $P_{READ-DISTURB} < 3 \times 10^{-12}$ over the entire range of NFET width and the optimum read current direction is to have parallelizing read.

Observe that the decision failure probability is strongly dependent on the NFET width when the NFET width is small. The decision failure probability then reaches a minimum and increases slightly with increasing NFET width. This trend may be explained by the effects of NFET channel resistance and MTJ resistance on the bit-cell TMR. For small NFET widths during read operation, the total bit-cell resistance is dominated by the NFET channel resistance which has little dependence on the MTJ configuration. Hence, the difference in bit-cell resistance when the MTJ is in parallel and when the MTJ is in anti-parallel is small. Under process variations, it becomes difficult to differentiate the resistances, resulting in higher decision failure probability. However, the total bit-cell resistance is dominated by the MTJ resistance when the access NFET becomes large enough. At larger NFET widths, the bit-cell TMR is

Table 4.2.
Parameters for Optimized STT-MRAM Bit-cells

	45 nm Bulk CMOS	45 nm SOI	16 nm PTM
Bit-cell Configuration	SC	SC	SC
Read Current Direction	Anti-parallelizing	Parallelizing	Parallelizing
NFET Width	1829 nm	1005 nm	463 nm
Failure Probability	$\sim 3.4 \times 10^{-6}$	$\sim 3.4 \times 10^{-6}$	$\sim 1.177 \times 10^{-7}$
Dominant Failure	Read-*decision*	Read-*decision*	Read-*decision*
V_{READ}	0.1 V	0.1 V	0.1 V
$I_{REF-OPT}$	27.83 μA	27.29 μA	23.02 μA

determined almost exclusively by the TMR of the MTJ. When the MTJ resistance dominates, the voltage dropped across the bit-cell is mostly dropped across the MTJ. Increasing NFET width decreases the NFET channel resistance and the overall bit-cell resistance while increasing V_{MTJ}, which reduces the TMR of the MTJ and the bit-cell. Since the voltage across the bit-cell is small during read operations, the increase in V_{MTJ} with increasing NFET width is very small. Hence, decision failure probability increases slightly with increasing NFET width.

Parameters for the optimum bit-cells are summarized in Table 4.2. The failure probability is reduced by more than an order of magnitude when MTJs are scaled. The likely reason is that MTJ resistances are significantly larger as compared to the NFET resistance [see Fig. 4.2(a)]. Note that in Fig. 4.2(b), TMR of the scaled MTJ is lower. Since the bit-cells using the MTJ characteristics assumed here are dominated by decision failure and since TMR expresses the relative change in MTJ resistance, MTJs with large relative and absolute resistance difference between P and AP states are needed to improve decision failure probability.

4.4 Summary

In this chapter, an optimization methodology for 1T-1MTJ STT-MRAM bit-cells is proposed. The application of the proposed optimization methodology to 1T-1MTJ STT-MRAM bit-cells in 45 nm bulk CMOS, 45 nm SOI and 16 nm PTM technologies is also studied. The MTJ characteristics used in this study were generated using the simulation framework proposed in Chapter 2, which was calibrated to experimentally reported data. The optimization methodology proposed in this chapter successfully optimized the bit-cell configuration as well as the NFET size. Furthermore, it is observed that resistance distinguishability in 1T-1MTJ STT-MRAM bit-cells depends strongly on the relative as well as the absolute resistance difference between MTJs in P state and MTJs in AP state.

5. ASSIST TECHNIQUES FOR FAILURE MITIGATION IN 1T-1MTJ STT-MRAM

Process variations may cause failures in STT-MRAM as was shown in Chapter 4. The analysis and optimization methodology proposed in Chapter 4 was applied to 1T-1R STT-MRAM bit-cells with read failures that are dominated by: 1) disturb failure and 2) decision failure. The results are summarized in Fig. 5.1, and the common trends observed are: 1) when the access transistor (ATx) is sized larger than the optimum width, read failure dominates; 2) write failure dominates when ATx is smaller than the optimum width; and 3) if write failure can be mitigated to shift the curve to the left, the optimum width of Tx may be reduced and the failure probability of the bit-cell may possibly reduce as well. Thus, to enable higher integration density and reduce the optimum area of 1T-1R bit-cells, techniques for reducing write failures need to be developed.

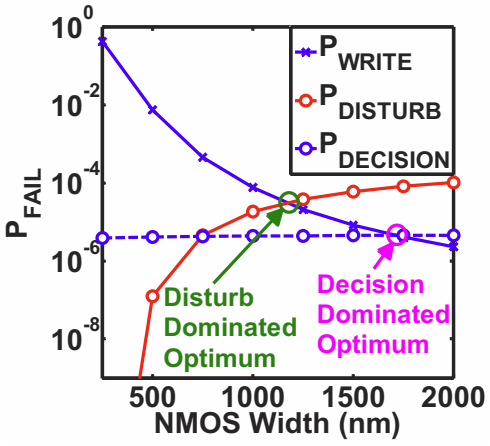


Fig. 5.1. Optimization results of disturb-failure-dominant and decision-failure-dominant bit-cells using the methodology from Chapter 4.

One method of reducing write failures is to reduce the critical switching current, I_C, of the MTJ. At the device level, significant research has been done to reduce I_C to lower write energy and write delay [28,36,44,45]. However, device-level techniques may require costly changes to the fabrication process. Circuit-level techniques that require minimal changes in the fabrication process are thus preferred for reducing I_C. One such assist technique using an applied external magnetic field was proposed in [57] and [58]. The authors of [57] suggest that the magnetic field be generated using the same current that flows through the MTJ. The magnetic field generated is small (< 0.5 Oe) and possibly insufficient to reduce I_C. Instead, additional structures for generating the assist magnetic field may be used, and will be presented later in Section 5.1.4. The main contribution of this chapter is the proposal of assist techniques for mitigating failures in 1T-1R STT-MRAM bit-cells. Specifically, several circuit-level write-assist techniques are developed. The techniques developed in this chapter may be used in conjunction with the optimization methodology proposed in Chapter 4 to yield smaller 1T-1R STT-MRAM bit-cells that are optimized for failures. The novelty of the approach in this dissertation compared to prior work, such as those in [59] and [60], is that no assumptions are placed on the distributions in bit-cell currents. Instead, distributions for bit-cell parameters such as MTJ cross-sectional area (A_{MTJ}), MTJ oxide thickness (t_{MgO}), and access transistor parameters are assumed because the bit-cell currents may not be normally distributed. Furthermore, the approaches proposed in [59] and [60] are architecture-level solutions. The approach in this chapter is a bit-cell-level and circuit-level solution complementing prior work.

5.1 Write Assist Techniques

As discussed earlier, the optimum ATx width may be reduced by mitigating write failures, possibly reducing bit-cell failure probability as well. Thus, four write failure mitigation techniques that reduce the optimum bit-cell size while maintaining bit-cell

performance are explored. The four write-failure mitigation techniques are *word-line voltage boosting*, *transistor body biasing*, *write voltage boosting*, and *external applied magnetic field*. The main idea behind the techniques that manipulate the voltages on the control lines (bit, source, and word lines, denoted BL, SL, and WL, respectively) or the body terminal of ATx is that current flowing through the MTJ during write operations may be increased. Alternatively, I_C needed to switch the MTJ may be reduced using an external applied magnetic field, as will be discussed in Section 5.1.4.

5.1.1 Word-line voltage boosting

In some transistor technologies, the transistor gate voltage may be boosted such that $V_{GS} > V_{DD}$. For bit-cells implemented with such transistor technologies, the word-line voltage (V_{WL}) may be boosted during write operations to lower the transistor resistance (R_{Tx}) and allow more current to flow through the MTJ. This is illustrated by the example load line in Fig. 5.2(a). In the SC bit-cell, the MTJ in the anti-parallel (AP) configuration may have such a large resistance (R_{MTJ}) that the current flowing through the MTJ (I_{MTJ}) falls below I_C. By boosting the word-line voltage, R_{Tx} is reduced drastically and, as a result, I_{MTJ} can rise above I_C . The analysis in Section 5.2 assumes a boosted V_{WL} of 1.3 V for write operations and $V_{WL} = 1.0$ V for read operations for the word-line voltage boosting assist technique. Since write operations in memory occur infrequently, boosting V_{WL} during write may have little impact on the reliability of the transistor. Also, note that in conventional 6T SRAMs, unselected cells in a row need to be placed in a pseudo-read condition and hence, a boosted V_{WL} may lead to disturb failures in the unselected cells during write operations. This is not the case in STT-MRAMs since the BL and SL in unselected columns may be discharged to GND to save power.

5.1.2 Write voltage boosting

Writing data into STT-MRAM bit-cells requires the application of voltages on BL, SL, and WL. WL controls the gate of ATx as well as I_{MTJ}. When V_{WL} is $V_{DD} = 1.0$ V, BL and SL voltages determine I_{MTJ}. Fig. 5.2(b) shows the D.C. load line of the bit-cell when the voltage on BL (V_{BL}) is V_{DD} and the voltage on SL (V_{SL}) is GND. The MTJ is in the AP configuration and cannot be written in the write cycle because $I_{MTJ} < I_C$. However, I_{MTJ} may be increased by increasing V_{BL} beyond V_{DD}, as shown in Fig. 5.2(b) by the dashed load line. When V_{BL} is increased, I_{MTJ} becomes

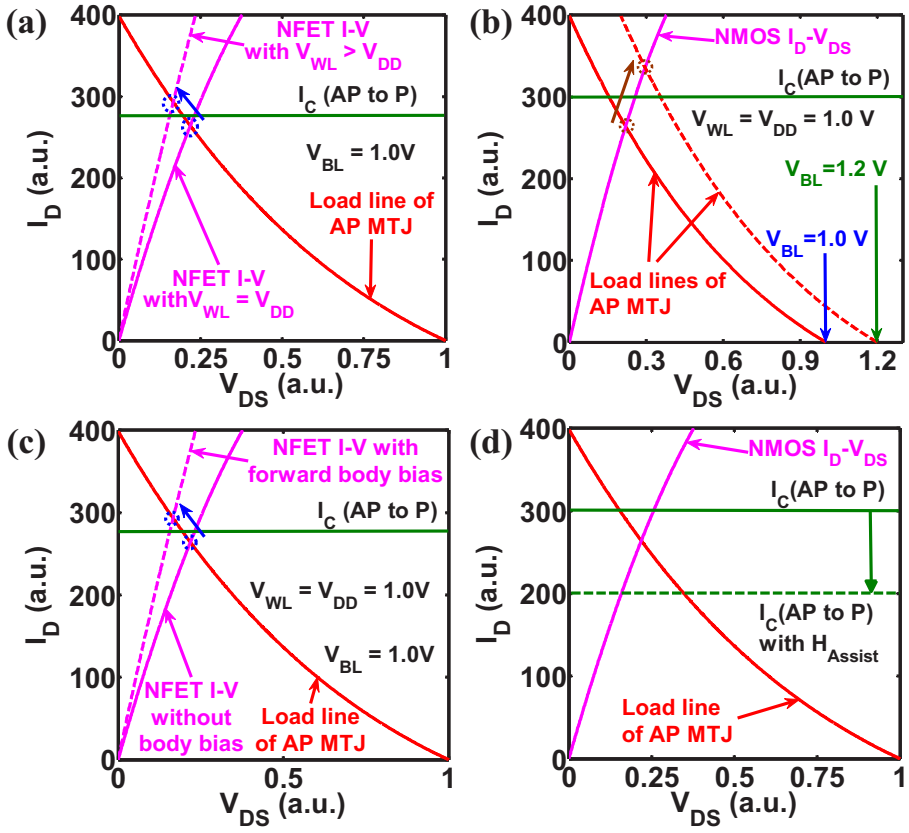


Fig. 5.2. These load lines illustrate how write failures are mitigated by (a) word-line voltage boosting, (b) write voltage boosting, (c) ATx body biasing, and (d) applied magnetic field assist. The transistor I_D-V_{DS} is shifted by word-line voltage boosting and ATx body biasing. The MTJ load line is shifted by write voltage boosting. The critical switching current (I_C) is shifted by applied magnetic field assist.

larger than I_C, and thus the MTJ can be successfully written during the write cycle. In order to implement the boosted write voltage, an additional voltage plane may be required along with voltage level converters in BL and SL drivers. Furthermore, a higher write voltage increases the electrical stress on the MgO barrier and may lead to reliability issues, which is beyond the scope of this dissertation. In Section 5.2, the write voltage ($V_{WRITE} = |V_{BL} - V_{SL}|$) is assumed to be 1.3 V at $V_{WL} = 1.0$ V for the write voltage boosting assist technique.

5.1.3 Access transistor body biasing

The bit-cell current may be increased without increasing the width of ATx by lowering the threshold voltage (V_T) of ATx, as shown in Fig. 5.2(c). A circuit-level technique to lower the V_T of ATx is to apply a voltage to the body of ATx (V_{BODY}). Assuming that inter-die variation is the dominant component of variation, a single bias to ATx body may be sufficient in improving the failure probability of the bit-cells on the die. Forward biasing the ATx body may increase the leakage currents in the cell, resulting in increased power dissipation. However, the increase in power may be insignificant because the array can be powered down during standby, and for small V_{READ}, the increase in junction leakage may be insignificant. Note that the only area penalty comes from circuitry for generating the body bias and not from bit-cells. In Section 5.2, $V_{BODY} = +0.3$ V body bias to all ATx in the array is assumed for the ATx body biasing assist technique, and $V_{WL} = 1.0$ V for read and write operations.

5.1.4 External applied magnetic field assist

It was shown in [45] that variations in switching delay of an MTJ are due to thermal fluctuations that cause the magnetization of the free layer (FL) to become non-collinear with the magnetization of pinned layer (PL). When FL and PL are exactly collinear, no spin-transfer torque can be generated and spin-transfer torque switching is impossible. Thermal fluctuations slightly perturb the FL magnetization

such that FL and PL are non-collinear, and spin-transfer torque may be generated when a current flows through the MTJ. Once the current starts flowing through the MTJ, the spin-transfer torque starts moving the FL magnetization away from the easy axis of the PL. When the angle between the magnetization of FL and the easy axis of the PL becomes large enough, the spin-transfer torque can overcome the anisotropies of the FL and switch the FL magnetization. However, an incubation period is required from the start of current flow before the spin torque exerted on the FL becomes large enough to switch the FL magnetization.

An alternative method to that proposed in [45] to reduce the incubation period is to apply a small magnetic field ($\vec{H}_{Assist}$) that tilts the FL magnetization towards its hard axis, as proposed in [57] and [58]. Compared to ATx body biasing, word-line voltage boosting, and write voltage boosting techniques, where bit-cell currents are increased and may affect the reliability of the MTJ, the applied magnetic field effectively reduces the I_C of the MTJ as shown in Fig. 5.2(d). However, the $\vec{H}_{Assist}$ required depends on the critical field of the FL ($\vec{H}_C$)

$$|\vec{H}_C| = \frac{2E_A}{M_S V} \tag{5.1}$$

where E_A, M_S, and V are the activation energy, saturation magnetization, and volume of the FL, respectively.

Analysis of the effects of a hard axis field on the switching delay of MTJs was done in [58, 61, 62]. It was found in the analysis done in this dissertation that I_C was not significantly reduced if $\vec{H}_{Assist}$ was turned on for the entire period when the bit-cell was being written. This is consistent with results reported in [62]. The reason is that the effect of the hard axis field is different during switching of the FL magnetization. The cause of this difference is the precessional nature of FL switching. When spin torque is turning the FL magnetization away from $\vec{H}_{Assist}$, $\vec{H}_{Assist}$ impedes spin torque. On the other hand, when spin torque is turning the FL magnetization towards $\vec{H}_{Assist}$, $\vec{H}_{Assist}$ aids spin torque. As a result, the overall switching delay is not significantly reduced. However, a significant reduction in I_C may be achieved if $\vec{H}_{Assist}$ is pulsed before spin-torque current starts flowing. In

the analysis done in [62], $\vec{H}_{Assist}$ was turned on first and the FL magnetization was allowed to settle before $\vec{H}_{Assist}$ is turned off and then injecting spin-torque current. This method achieved significant reduction in I_C for sufficiently large $\vec{H}_{Assist}$. The technique proposed in this chapter differs from that proposed in [62] in that $\vec{H}_{Assist}$ is turned on for a fixed period, regardless of whether the FL magnetization has settled. Compared to this scheme, the scheme proposed in [62] trades off settling time and the spin-torque current pulse width. If the time for FL magnetization to settle at its equilibrium is large, the pulse width for spin-torque current must be small so that the total write delay is constant. The total write delay is 40 ns in this chapter. If $\vec{H}_{Assist}$ is pulsed for 10 ns and the FL magnetization is allowed to settle, the spin-torque current pulse must be 30 ns to meet the write delay target. Because of the inverse exponential dependence of I_C on switching delay, the reduction in I_C due to $\vec{H}_{Assist}$ may be cancelled by the increase in I_C due to reduction of spin-torque current pulse width from 40 to 30 ns. Hence, the I_C reduction at larger $\vec{H}_{Assist}$ shown here is not as much as reported in [62]. Furthermore, large $\vec{H}_{Assist}$ may not necessarily improve the write failure probability at iso-write cycle time. When $\vec{H}_{Assist}$ is just turned on, the FL magnetization experiences a significant disturbance. As shown in [62], the

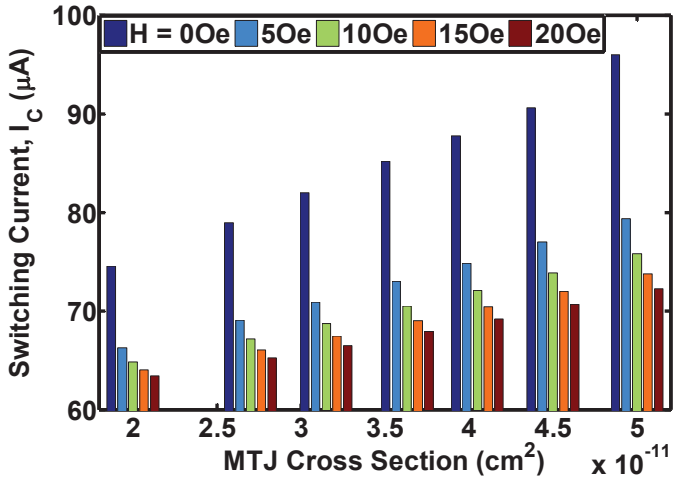


Fig. 5.3. Iso-E_A switching current for AP to P with different applied magnetic fields and MTJ cross-sectional area.

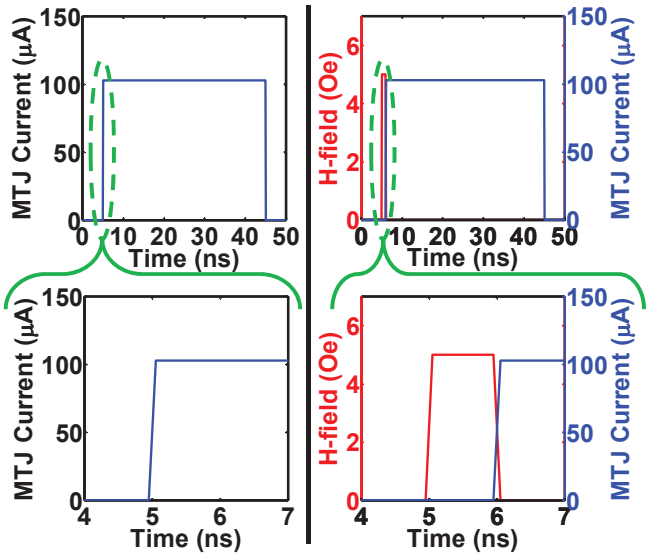


Fig. 5.4. Timing diagram of assist magnetic field (below) and the current pulse that flows through the MTJ with (below) and without (top) assist magnetic field.

response of the FL magnetization may overshoot the final state when $\vec{H}_{Assist}$ is just turned on. When this occurs, significant reduction in I_C may be observed even in the presence of thermal fluctuations. However, this occurrence depends on the thermal fluctuation as well as the initial magnetization of FL. The variation in switching delay is increased if FL magnetization is not allowed to settle. Hence, when a large $\vec{H}_{Assist}$ is applied, a significant amount of time is required for FL magnetization to stabilize so as to maintain switching delay variation.

The amount of power consumed to generate $\vec{H}_{Assist}$ may also be significant. For example, 250 μA is needed to generate 5 Oe of magnetic field 100 nm away from a straight interconnect wire. However, this may be reduced by cladding the wire with a suitable material, as demonstrated in [63]. Also, if $\vec{H}_{Assist}$ is turned on for the entire duration of write as proposed in [58], the power consumption will be very large. Since $\vec{H}_{Assist}$ is only required to destabilize the initial magnetization of the FL during write, it does not need to be turned on for the entire duration of write. By reducing the amount of time $\vec{H}_{Assist}$ is turned on, the overall power consumption may be lower

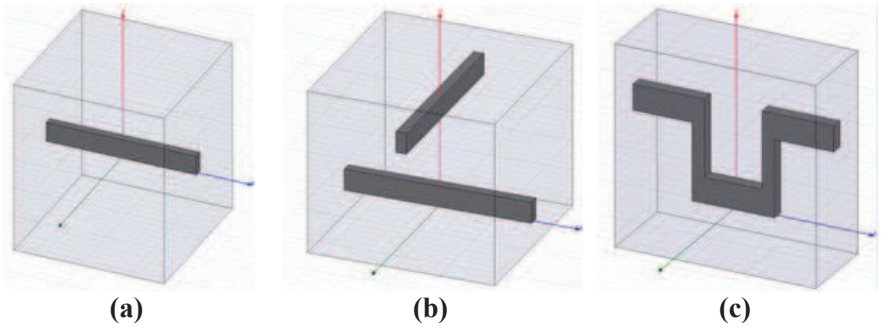


Fig. 5.5. Interconnect structures that can be used to generate assist magnetic field. The MTJ is situated along the vertical axis.

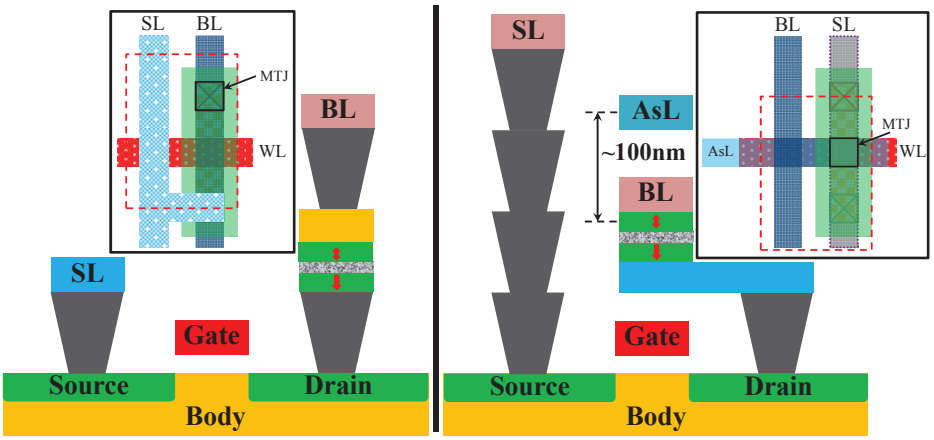


Fig. 5.6. Layouts of bit-cell structures (left) without magnetic field generating structure, and (right) with a long interconnect wire to generate magnetic field for assisting write (labeled "AsL"). (Inset) Top-down view of cells with the MTJ (black boxes). The bit-cell area (red dashed boxes) is the same in both cases.

than that without $\vec{H}_{Assist}$. In our analysis, the $\vec{H}_{Assist}$ pulse is assumed to be on for 1 ns and I_{MTJ} to flow for 39 ns immediately after the $\vec{H}_{Assist}$ pulse is turned off. The graph in Fig. 5.3 shows the iso-activation energy (iso-E_A) reduction in I_C when switching from AP to P for different strengths of $\vec{H}_{Assist}$ and for different A_{MTJ}. The timing diagrams for the current pulses through the MTJ with and without $\vec{H}_{Assist}$ are shown in Fig. 5.4. When $\vec{H}_{Assist} = 0$ Oe, I_{MTJ} flows for 40 ns. Generally, the largest reduction in I_C occurs when $\vec{H}_{Assist}$ is small.

On-chip structures for generating magnetic fields

A method of generating $\vec{H}_{Assist}$ on-chip was proposed in [57]. However, other structures may be used to generate $\vec{H}_{Assist}$. In this dissertation, three structures (shown in Fig. 5.5) that may be incorporated into STT-MRAM arrays to generate $\vec{H}_{Assist}$ on-chip are proposed. The MTJ is situated along the vertical axis (red line) in Fig. 5.5. The need for additional structures to generate $\vec{H}_{Assist}$ may result in an area overhead compared to the standard STT-MRAM bit-cell. In this dissertation, bit-cells incorporating the structure shown in Fig. 5.5(a) are analyzed. The interconnect wire runs parallel to WL and the FL of the MTJ sits 100 nm below the wire, as shown in Fig. 5.6. The wire also has no cladding that can reduce the current required to generate $\vec{H}_{Assist}$. In the proposed layout shown in Fig. 5.6, $\vec{H}_{Assist}$ acting on the nearest neighbor MTJ is less than half of $\vec{H}_{Assist}$ acting on selected MTJs. Since $\vec{H}_{Assist} = 5$ Oe is very small compared to the critical field (~ 900 Oe) of the FL, the disturbance on unselected neighboring MTJs is negligible.

5.2 Comparison of Write Assist Techniques

Bit-cells implemented with each failure mitigation technique were optimized at iso-delay to compare the effectiveness of individual failure mitigation technique. Table 5.1 lists each of the cases analyzed, their associated parameters, and results of optimization corresponding to each case. Write power is calculated by averaging the average energy per write operation over the write cycle period ($t_{write} = 40$ ns). The average write energy (AWE) is computed as

$$\text{AWE} = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{E_{i,j}}{4} \tag{5.2}$$

where $E_{i,j}$ is the energy to write data 'j' into a bit-cell storing data 'i'. In the bit-cells analyzed in this work, only $E_{0,0}$, $E_{0,1}$, $E_{1,0}$, and $E_{1,1}$ are available. The write power is then calculated as

$$\text{Write Power} = \frac{\text{AWE}}{t_{write}} \tag{5.3}$$

With the exception of transistor body biasing, all other failure mitigation techniques do not affect read failures. Even though disturb failures increased they remain negligible compared to decision and write failures. Thus, the optimum ATx width is still determined by write and decision failures. Note that read power increase is negligible, even when ATx body biasing technique is applied. This is because the ATx width is smaller and V_{READ} is small enough that junction leakage is not significantly increased compared to the read currents through the MTJ. Also, improvement in decision failure was observed only for ATx width below 400 nm. Overall, ATx body biasing shifts the ATx width versus decision failure curve towards the left. Also, increasing ATx body bias increases the minimum achievable decision failure. The results show that the optimum decision failure probability occurred at ATx width of 908 nm with no body bias, compared to 876 nm for $V_{BODY} = +0.3$ V. However, the decision failure probability increased from 3.3718×10^{-6} to 3.3723×10^{-6}. The cause is that body biasing increased the ATx drive strength and allowed more current to flow through the bit-cell for reading; nominal read currents for AP and P configurations increased from 17.58 μA and 43.12 μA to 17.60 μA and 43.21 μA, respectively. Hence, $I_{REF-OPT}$ increased from 26.75 μA to 26.79 μA after ATx body biasing was applied. At increased read currents, TMR of the MTJ is reduced, even though sensing margins increased from 9.17 μA to 9.19 μA. By approximating the read current as

$$I_{READ} = \frac{V_{READ}}{R_{Tx} + R_{MTJ}} \tag{5.4}$$

where R_{Tx} is the ATx channel resistance (relatively constant for small V_{READ}), an increase in I_{READ} leads to smaller nominal R_{MTJ} and increases sensitivity to variations in t_{MgO} and A_{MTJ}. Thus, sensing margins alone may not be able to accurately gauge the sensing failure rates in STT-MRAMs. Furthermore, additional failure mitigation techniques [30, 31], which are beyond the scope of this dissertation, may be required to further improve array yield of STT-MRAMs.

Table 5.2 and Table 5.3 lists the results of iso-ATx width and iso-failure probability comparisons of the techniques listed in Table 5.1, respectively. When individual assist techniques are compared, word line voltage boosting technique achieved the

Table 5.1.
Simulation Parameters and Optimization Results for 1T-1R STT-MRAM Bit-cells Analyzed

Assist Technique Applied	Technique Parameters	Optimum Tx Width (nm)	$I_{REF-OPT}$ (μA)	Sensing Margin (μA)	Calculated P_{FAIL}	Relative Write Power	Relative Bit-cell Area
A. No Assist Technique	V_{WL}=1.0V V_{WRITE}=1.0V V_{BODY}=0.0V	1829	27.97	9.87	3.39×10^{-6}	1.0	1.0
B. Tx Body Biasing	V_{BODY}=+0.3V	1333	27.59	9.65	3.38×10^{-6}	0.975	0.744
C. V_{WRITE} Boosting	V_{WRITE}=1.3V	1180	27.30	9.48	3.38×10^{-6}	1.036	0.653
D. Applied External Magnetic Field	1ns pulse, 5Oe, Fig. 5.5(a) structure	908	26.75	9.17	3.372×10^{-6}	0.874	0.497
E. Word-line Voltage Boosting	V_{WL}=1.3V	908	26.75	9.17	3.372×10^{-6}	1.04	0.497
F. Technique B + Technique C	–	942	26.95	9.28	3.373×10^{-6}	1.01	0.515
G. Technique C + Technique D	–	908	26.75	9.17	3.372×10^{-6}	0.973	0.497
H. Technique D + Technique E	–	908	26.75	9.17	3.372×10^{-6}	1.067	0.497
I. Technique C + Technique E	V_{WL} = 1.3V V_{WRITE}=1.3V	908	26.75	9.17	3.372×10^{-6}	1.194	0.497
J. Technique C + Technique E	V_{WL} = 1.1V V_{WRITE}=1.1V	908	26.75	9.17	3.372×10^{-6}	1.05	0.497

Table 5.2.
Write Failure Probability of Table 5.1 Techniques at 500nm Transistor Width

A	B	C	D	E	F	G	H	I	J
7.49×10^{-3}	2.32×10^{-3}	1.25×10^{-3}	1.61×10^{-4}	2.58×10^{-8}	4.21×10^{-4}	1.65×10^{-5}	3.41×10^{-11}	1.33×10^{-9}	8.29×10^{-5}

Table 5.3.
Transistor Width of Table 5.1 Techniques at 1×10^{-4} Failure Probability

A	B	C	D	E	F	G	H	I	J
957 nm	766 nm	703 nm	528 nm	239 nm	596 nm	415 nm	169 nm	201 nm	740 nm

best reduction in failure probability and ATx width at iso-ATx width and at iso-failure probability, respectively, followed by the applied external magnetic field technique. The structure assumed for generating $\vec{H}_{Assist}$ is a straight line interconnect [see Fig. 5.5(a)] and hence, the layout of the cell is similar to that of field-switched MRAM and the interconnect may run above or below the MTJ as required to reduce the footprint of the bit-cell [64]. Furthermore, ATx is much larger than minimum size to provide sufficient write current through the MTJ thus, enlarging the bit-cell footprint. The additional area may then be used for the interconnect structures for generating $\vec{H}_{Assist}$. Note that even though currents as large as 250 μA are required to generate $|\vec{H}_{Assist}| = 5$ Oe of assist hard axis field, the total write energy is reduced because of the short pulse of $\vec{H}_{Assist}$ (1 ns) and reduction in I_C. The energy for generating $\vec{H}_{Assist}$ field is 2.5 fJ if the current is directly drawn from the V_{DD} supply (corresponding to 62.5 nW for $t_{write} = 40$ ns).

Schemes employing a combination of write failure mitigation techniques were also analyzed and the results are listed in Table 5.1. Note that minimizing P_{FAIL} was used as the optimization criteria. The minimum achievable P_{R-DEC} occurred at 908 nm and at 876 nm for $V_{BODY} = 0$ V and $V_{BODY} = +0.3$ V, respectively. Hence, the optimum widths cannot be lower when decision failure is the dominant failure below 908 nm and 876 nm, respectively. Since decision failure does not vary significantly for a range of widths below the optimum width (as seen in Fig. 4.7), architecture level failure mitigation techniques, such as those analyzed in [65], may be used to mitigate read failures in conjunction with our write failure mitigation techniques to achieve much smaller ATx width and hence, smaller array area. Hence, using combinations of read and write failure mitigation techniques, the total array power consumption and the data storage density may be reduced and increased, respectively.

5.3 Summary

Four write failure mitigation techniques – access transistor body biasing, write voltage boosting, word-line voltage boosting and external applied magnetic field technique – were developed and analyzed in this chapter. Using the optimization technique and bit-cell failure estimation methodology proposed in Chapter 4, bit-cells designed with and without assists were optimized and compared at iso-write delay. For the MTJ used in the analysis, external applied magnetic field assist was the most efficient among the four techniques. Bit-cells implemented with external applied magnetic field generated using a long current carrying wire achieved reduction in optimum access transistor width and write power consumption.

6. ALTERNATIVE STORAGE ELEMENTS FOR STT-MRAM

In the previous chapters, the design and optimization of standard spin-transfer torque magnetic RAM (STT-MRAM) have been discussed. Chapter 3 discussed the three main failure mechanisms in STT-MRAMs that have been explored in this dissertation – *write failure*, *read-disturb failure*, and *read-decision failure*. Then, circuit-level failure mitigation techniques were proposed and evaluated in Chapter 5. These techniques are preferred because they do not require changes to the fabrication process of the magnetic tunnel junction (MTJ), which is the storage device in STT-MRAM. However, as will be discussed later in this chapter, improvements in the characteristics of the storage device are required to fully exploit the benefits of STT-MRAM in memory systems.

The previous chapters showed that the main design issue that severely limits the minimum cell size of STT-MRAM for high-performance on-chip cache applications is the large critical write currents (I_C) required to program the MTJ. Thus, write failures are mitigated by ensuring that the access transistor (ATx) in the bit-cell is large enough to allow sufficient write current to flow through the bit-cell during write operations. In order to reduce the required ATx size and thus, reduce the bit-cell area to increase memory density of STT-MRAM, write failure mitigation techniques were proposed and evaluated in Chapter 5. Although the techniques successfully reduced the bit-cell area, it was shown in Fig. 4.7 that the lowest failure probability (P_{FAIL}) is limited by read-decision failure ($P_{READ-DECISION}$). Improvements in the distinguishability between the stored MTJ states may be achieved either by reducing the variations in the resistance of the MTJ or by increasing the Tunneling Magnetoresistance Ratio (TMR) of the MTJ. Both are improvements in the characteristics

of the MTJ and hence, device-level design techniques are required to exploit the full potential of STT-MRAM.

In this chapter, device-level design techniques to improve STT-MRAM for high-performance on-chip cache applications are explored. The first device that is evaluated is a multi-ferroic tunnel junction (MFTJ) in which the MgO tunnel oxide in the MTJ is replaced with a ferroelectric tunnel barrier. The improvements obtained in using an MFTJ are presented, followed by a discussion of the inherent limitations as a result of the two-terminal nature of the MTJ. This is then followed by a short discussion of some multi-terminal MTJ structures that have been proposed in the literature to mitigate the design issues arising from the limitations of using two-terminal MTJ as the storage device. However, since the devices proposed in the literature do not completely overcome the design issues, which will be discussed later, an alternate MTJ structure consisting of complementary polarizers (the CPMTJ) is proposed in Section 6.2.1. Analysis of the proposed CPMTJ, which will be presented later in this chapter, shows that it is able to solve the design issues in STT-MRAM based on the two-terminal MTJ, leading to significant improvements in STT-MRAM performance.

6.1 The Multi-ferroic Tunnel Junction

As mentioned earlier, sensing failures may severely limit the bit-cell area and failure probability in 1T-1MTJ STT-MRAM. Since it will be very challenging to significantly reduce process variations in the MTJ, enhancing the TMR of the MTJ may improve the sensing failure probability of STT-MRAM bit-cells. Replacing the tunnel barrier in an MTJ with a ferroelectric tunnel barrier (FTB) allows modulation of the tunneling conductance through the tunneling electroresistance (TER) effect [66], which may be used to enhance the TMR of the tunnel junction (TJ) and improve sensing failures in STT-MRAM memory cells.

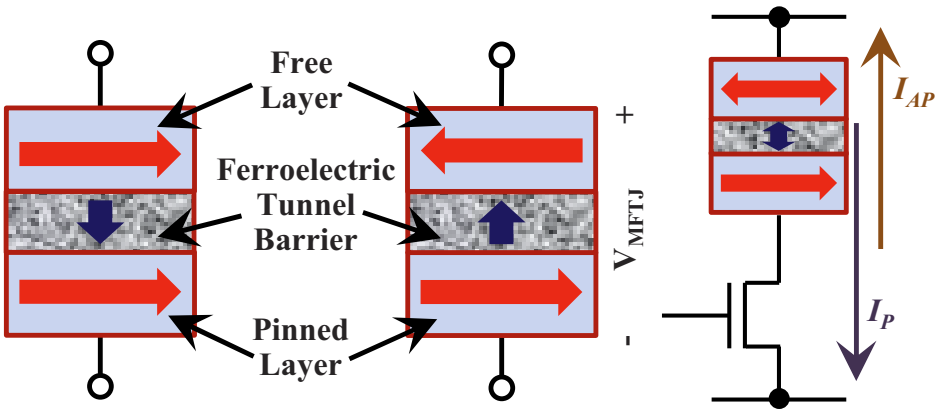


Fig. 6.1. The MFTJ structure consists of two ferromagnetic (FM) layers (blue with red arrows) sandwiching a thin ferroelectric layer (gray with dark blue arrows). The arrows denote the magnetization and electric polarization of the ferromagnetic and ferroelectric layers, respectively. In-plane anisotropy (IMA) FM layers are shown for illustration. The two memory states available are shown. (Right) The circuit schematic of the MFTJ based STT-MRAM memory cell with PL on the bottom. I_{AP} and I_P denote the current directions for anti-parallelizing and parallelizing the FM layers, respectively.

6.1.1 The MFTJ structure

The structure of the multi-ferroic tunnel junction (MFTJ, shown in Fig. 6.1) consists of two ferromagnetic electrodes sandwiching an FTB. Ferromagnetic configuration of the MFTJ is switched using spin-transfer torque like in MTJ-based STT-MRAM. The current directions for anti-parallelizing (I_{AP}) and for parallelizing (I_P) the FL are shown in Fig. 6.1. Since the FTB is very thin, the electric field in the tunnel barrier during write operations may be sufficient to switch the FTB polarization when current is being passed through MFTJ to switch its FL magnetization. Hence, two configurations of ferroic properties exist in the structure as shown in Fig. 6.1. The remnant polarization in the FTB and non-zero screening lengths in the ferromagnetic electrodes result in a small *TER* effect as illustrated by the band diagrams in Fig. 6.2. The effective potential along the transport direction of the MFTJ is such that the barrier height is larger when FTB polarization points toward the electrode with the

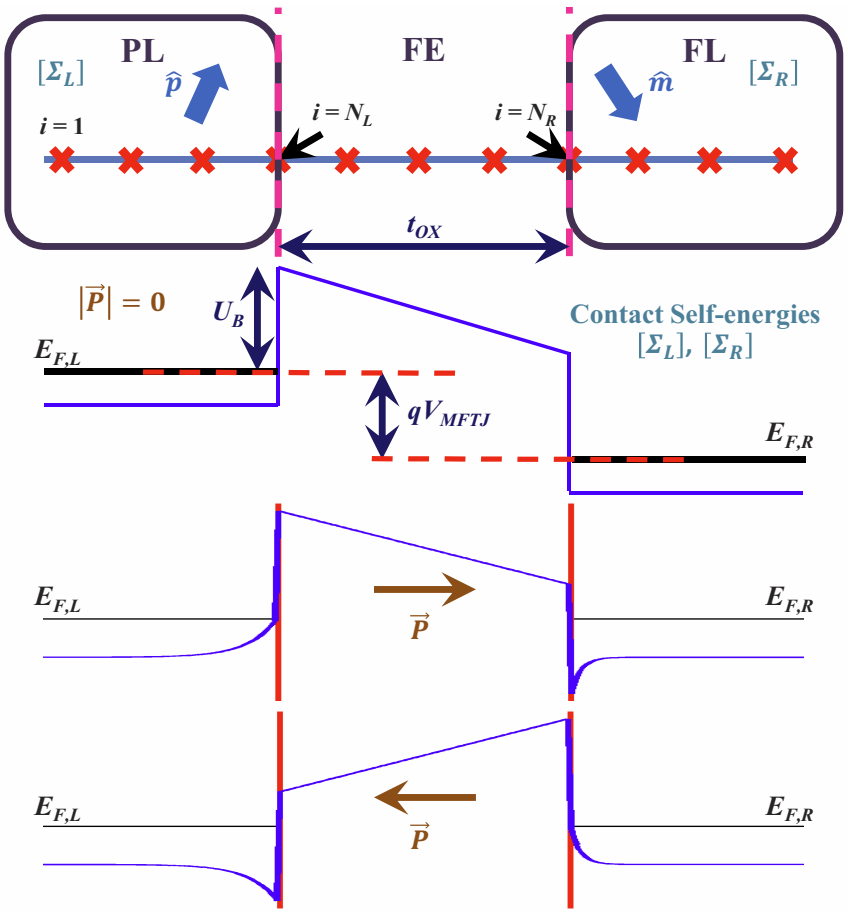


Fig. 6.2. Conceptual description of the MFTJ in the NEGF framework, where each cross represents a lattice point. The potential profile across the MFTJ under different FE polarizations without spin splitting is also shown.

larger screening length. Although the *TER* effect is small when FTB is thin, it may be sufficient to enhance the *TMR* of the MFTJ and hence, reduce sensing failures in STT-MRAM.

6.1.2 MFTJ modeling

The MFTJ may be modeled just like an MTJ, except that the physics due to the ferroelectric polarization need to be included in the model for the MFTJ. The dynamics of the ferroelectric polarization is modeled using the *Landau-Khalatnikov*

(LK) model. The Non-Equilibrium Green's Function (NEGF) solver explained in Appendix A is also modified to account for effects in the ferromagnetic contacts of the MFTJ induced by the ferroelectric polarization of the FTB.

The NEGF model for the MFTJ

The $I-V$ of the MFTJ may be calculated using the NEGF approach described in Appendix A. The effect of ferroelectric polarization in the tunneling barrier is modeled by adding to $\mathcal{H}$ in Appendix A an extra potential, U_{FE}, written as

$$U_{FE(i,i)} = \begin{cases} \sigma_S \phi_{L,i} I, \text{ if } i \leq N_L \\ \sigma_S \phi_{R,i} I, \text{ if } i \leq N_R \\ \left(\frac{N_R - i}{N_R - N_L} - \frac{1}{2}\right) \sigma_S \left(\frac{\delta_L}{\epsilon_L} + \frac{\delta_R}{\epsilon_R}\right), \text{ otherwise} \end{cases} \tag{6.1}$$

δ_l and ϵ_l are the Thomas-Fermi screening length and relative permittivity of electrode l, respectively. Also,

$$\phi_{l,i} = \frac{\delta_l e - \frac{|N_l - i|}{\delta_l}}{\epsilon_l} \tag{6.2}$$

$$\sigma_S = \left|\vec{P}\right|, \text{ where } \left|\vec{P}\right| \text{ is positive if } \vec{P} \text{ is pointing left in Fig. 6.2} \tag{6.3}$$

Finally, the current density flowing through the MFTJ (J_{MFTJ}) can be calculated using Eq. A.15. However, J_{MFTJ} depends on the magnetization directions of the pinned layer, PL, and free layer, FL (given by $\widehat{m} \cdot \widehat{M}$, where $\widehat{m}$ and $\widehat{M}$ are the magnetization directions of the FL and PL, respectively), and on the polarization of the ferroelectric tunnel barrier ($\vec{P}$). In this model, the dependence of J_{MFTJ} on $\widehat{m} \cdot \widehat{M}$ and on $\vec{P}$ are decoupled. Hence, for a fixed $\vec{P}$, $J_{MFTJ}(\theta)$ $\left(\theta = cos^{-1}\left(\widehat{m} \cdot \widehat{M}\right)\right)$ may be calculated using

$$J_{MFTJ}(\theta) = J_P \; cos^2\left(\frac{\theta}{2}\right) + J_{AP} \; sin^2\left(\frac{\theta}{2}\right) \tag{6.4}$$

where $J_P = J_{MFTJ}(\theta = 0)$ and $J_{AP} = J_{MFTJ}(\theta = \pi)$. $J_{MFTJ}(\left|\vec{P}\right|)$ may then be written as

$$J_{MFTJ}\left(\left|\vec{P}\right|\right) = e^{c_1 \left|\vec{P}\right| + c_0} \tag{6.5}$$

where c_i are fitting parameters (different for positive and for negative $\left|\vec{P}\right|$) since $\vec{P}$ modulates the effective barrier height [66].

The Landau-Khalatnikov model

Dynamics of ferroelectric polarization is described by the Landau-Khalatnikov (LK) equation [67] as given by

$$\frac{\partial \vec{P}}{\partial t} = -a_0 \frac{\partial F\left(\vec{P}\right)}{\partial \vec{P}} \tag{6.6}$$

where $F(\vec{P})$ is the free energy functional of the ferroelectric material, and a_0 is a proportionality constant. $F(\vec{P})$ is written as

$$F\left(\vec{P}\right) = F_0\left(\vec{P}\right) + a_1 \vec{E} \cdot \vec{P} \tag{6.7}$$

where $F_0(\vec{P})$ describes the ferroelectric anisotropy, a_1 is a proportionality constant, and $\vec{E}$ is the external electric field applied across the ferroelectric.

SPICE compatible model for the MFTJ

The SPICE compatible dynamical MTJ model developed in this dissertation was presented in Chapter 2. It was modified to enable SPICE simulations to include ferroelectric dynamics. The components of the modified SPICE model are shown in Fig. 6.3. Note the inclusion of an additional Ordinary Differential Equation (ODE) solver block to model the LK equation, on top of the two ODE blocks used to model the LLG equation in spherical coordinates. The $I - V$ characteristics of the MFTJ returned by our NEGF solver are encapsulated as a compact model, and may also include $\overrightarrow{STT}$ calculated using Eq. A.16 in the NEGF solver. Alternatively, the model for $\overrightarrow{STT}$ proposed in [68], written as Eqs. C.2–C.8, may also be used. Each ODE solver block consists of a capacitor network as shown in Fig. 6.3, where each current source represents one term in the differential equation and capacitor voltages are $\vec{P}$

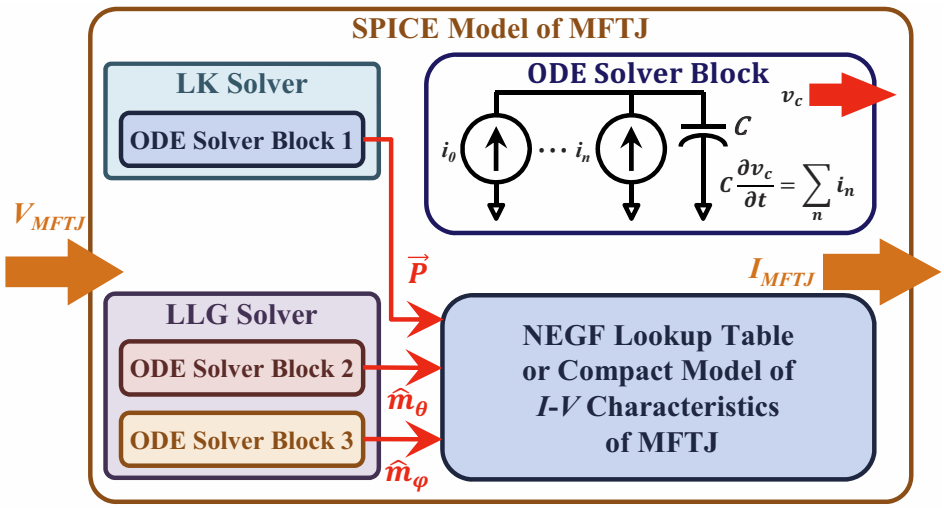


Fig. 6.3. Block diagram of the SPICE compatible MFTJ model proposed and developed in this dissertation.

and components of $\hat{m}$, in spherical coordinates in the LK and LLG block, respectively. $\vec{P}$ and $\hat{m}$ are used to calculate I_{MFTJ}, V_{MFTJ}, and $\overrightarrow{STT}$ during simulation.

6.1.3 Evaluation of MFTJ for STT-MRAM based high-performance on-chip cache

The MTJ characteristics used as the baseline for comparison are graphed in Fig. 2.6. Ferroelectric polarization was added to this MTJ to create an MFTJ for exploration. The ferroelectric polarization versus electric field hysteresis curve and the ferromagnetic parameters assumed for the MFTJ are shown in Fig. 6.4 (t_{OX} in the MFTJ case is the equivalent MTJ t_{OX}). $\vec{P}$ is assumed to be pointing along the direction of electron transport. Other device parameters are listed in Table 6.1.

TMR versus oxide voltage (voltage applied across the tunnel junctions) were calculated in our NEGF solver and plotted in Fig. 6.5, showing that the TMR of the MFTJ is 7.2% higher than that of the MTJ. However, the TMR of the MFTJ based STT-MRAM memory cell is only 4.7% higher than MTJ based STT-MRAM (assuming 900 nm wide ATx), implying that transistor resistance significantly affects

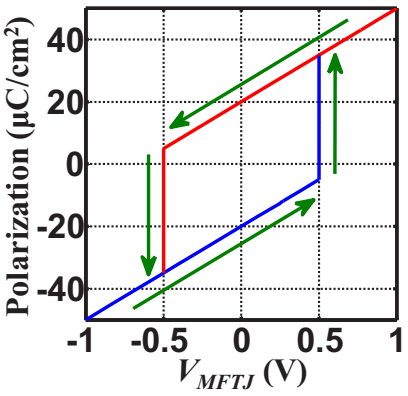


Fig. 6.4. Ferroelectric polarization vs. applied voltage curve of MFTJ.

Table 6.1.
Parameters of MFTJ Model

Gilbert Dampling, α	0.014
Gyromagnetic Ratio, γ	17.6 MHz/Oe
Free Layer Geometry	50 nm×50 nm×1.4nm
t_{WRITE}, I_{C0}	5 ns, 60 μA
ATx Technology	45 nm bulk CMOS
Retention Barrier (E_A)	$56k_BT$
t_{OX}, Read V_{DD}, V_{READ}	1.25 nm, 1.0 V, 0.3 V

TMR of the memory cell. Although the FTB enhanced the TMR of MFTJ based STT-MRAM, the overall resistance of the memory cell is also higher than that of MTJ based STT-MRAM. Consequently, read disturb current through the MFTJ based STT-MRAM is 0.3 μA lower than in MTJ based STT-MRAM. Read-disturb failures are thus lower in MFTJ based STT-MRAM than in MTJ based STT-MRAM. On the other hand, due to the larger resistance, MFTJ based STT-MRAM requires a write voltage of 0.973 V compared to 0.971 V in MTJ based STT-MRAM (considering 10% write margin, where write margin = $\frac{I_{WRITE}-I_{C0}}{I_{C0}} \times 100\%$).

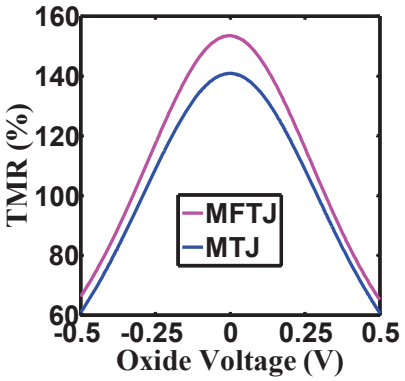


Fig. 6.5. Comparison of device TMR of MFTJ and MTJ.

As discussed in Chapter 1, it is extremely challenging to design robust STT-MRAM for on-chip cache applications due to conflicting design requirements. The problem is further compounded by the fact that the storage device is a two-terminal device, which limits the design choices available. On the other hand, multi-terminal MTJ structures provide an avenue to alleviate these design limitations. In the following sections, the discussion will focus on the design of STT-MRAM using these multi-terminal structures. A multi-terminal MTJ structure consisting of complementary polarized pinned layers will be proposed, and the later sections will show how the proposed structure enables STT-MRAM based cache to outperform 6T SRAM based cache.

6.2 Multi-terminal MTJs as STT-MRAM Storage Devices

It has been shown in the earlier sections that two-terminal MTJs for STT-MRAM requires the read and write current paths to be shared, which leads to severe design limitations. Although the two-terminal nature of the storage device allows for very small bit-cell footprint, the benefits are eroded if better STT-MRAM performance is required. Several multi-terminal MTJ structures have been proposed in the literature to mitigate the aforementioned design issues. Although multi-terminal MTJ structures require additional ATx in the bit-cell, the sizing requirements on the ATx

may be less stringent than that in STT-MRAM based on two-terminal MTJs and hence, STT-MRAM bit-cells using multi-terminal MTJ structures may have smaller footprint than STT-MRAM bit-cells based on two-terminal MTJs. A review of the multi-terminal MTJ structures proposed in the literature is presented in Appendix D. In this dissertation, a novel multi-terminal MTJ is proposed and evaluated for high-performance on-chip cache application.

6.2.1 The complementary polarizer MTJ structure

Fig. 6.6 shows the structure of the proposed complementary polarizer MTJ (CPMTJ) with perpendicular magnetic anisotropy (PMA) and its corresponding array organization. The design of CPMTJ based STT-MRAM, or CPSTT, is guided by the key insight that the parallelizing operation is preferred to reduce write power because I_C required to align magnetic layers is usually lower than that to anti-align magnetic layers [51]. The structure of the CPMTJ consists of two complementary polarized PL, and one FL sandwiching a tunneling oxide. Write operations in CPSTT occur by steering current through the bit-cell depending on the data being stored, as illustrated in Fig. 6.7, and hence CPSTT requires two ATx's (ATxL and ATxR). Although two ATx's are required, their sizing requirement is relaxed because there is no source degeneration in the write operation of CPSTT. The FL is connected to the bit-line (BL) while ATxL is connected to the left source-line (SLL) and ATxR is connected to the right source-line (SLR). Current flows from BL to SLL to write '0' (FL becomes parallel to the left PL, which is connected to SLL), whereas current flows from BL to SLR to write '1' (FL becomes parallel to the right PL, which is connected to SLR). Note that in CPSTT, the data is represented by the FL magnetization relative to two complementary PL magnetizations.

During read operations, the voltages of SLL, SLR, *SD* and *SDB* of the sense amplifier are first charged to V_{Pre} [see Fig. 6.8 and Fig. 6.9] by setting *RCLK*, *RDEN* and *REN* to V_{DD}. After *SD* and *SDB* are charged to V_{Pre}, *RCLK* is set to GND to

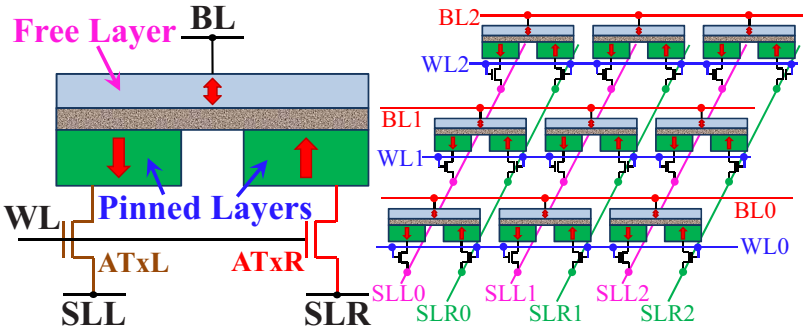


Fig. 6.6. (Left) Proposed Complementary Polarizer STT-MRAM structure (CPSTT), and (right) the organization of CPSTT memory array. Only three rows and three columns are shown to illustrate array organization.

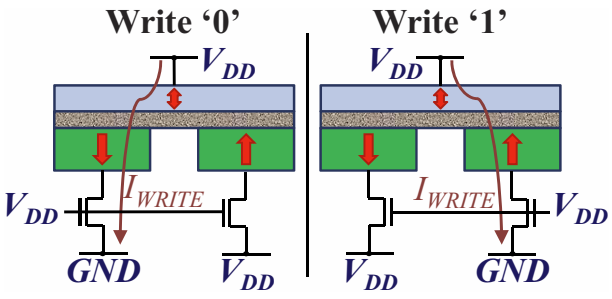


Fig. 6.7. Voltages across and currents flowing through our CPSTT bit-cell during write operations, and the physical representation of '0' and '1' states.

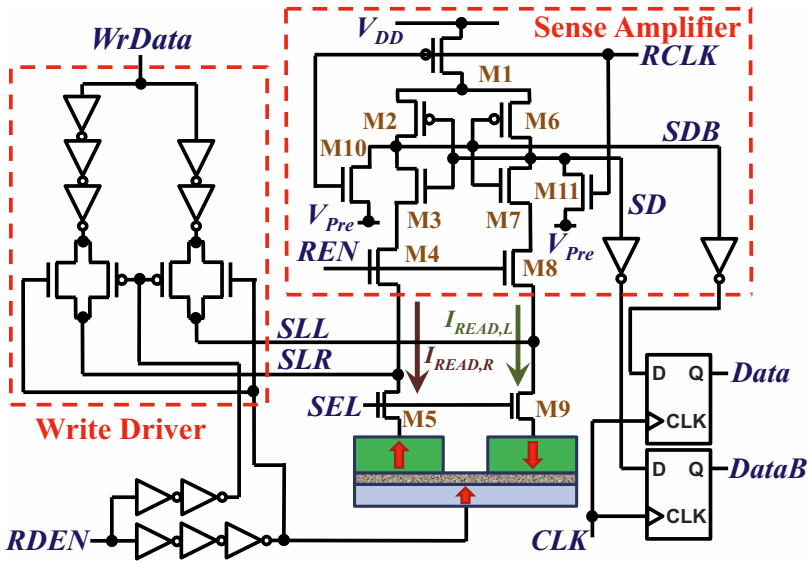


Fig. 6.8. Source-line (SL) and bit-line (BL) drivers, and latch based sense amplifier for CPSTT. Control circuitry for *SEL* (row decoders), *REN*, *RDEN*, *RCLK*, *CLK*, *WrData*, and column selection multiplexers are not shown.

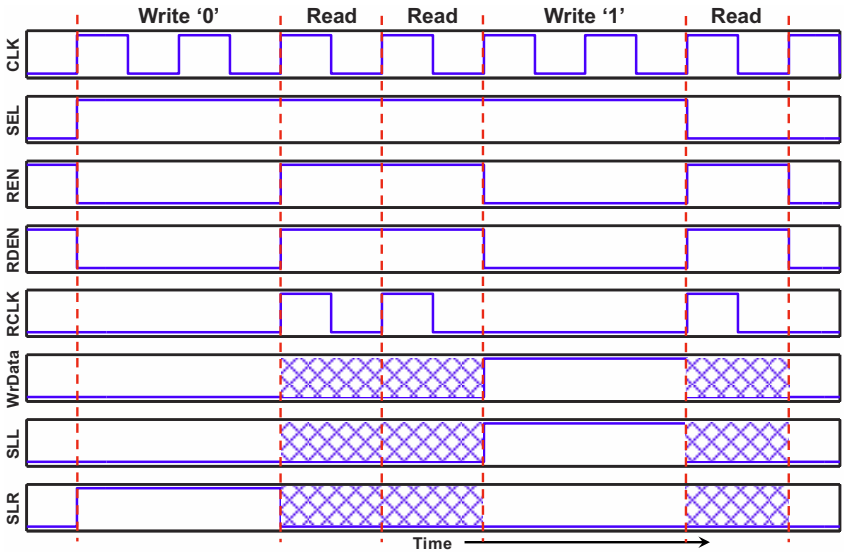


Fig. 6.9. Timing diagram of control signals *SEL*, *REN*, *RDEN*, *RCLK*, *WrData*, SLL, and SLR during write and during read operations, relative to the clock (*CLK*) signal. *WrData* is the data to be written during write operations, and $GND \leq (V_{SLL},\ V_{SLR}) \leq V_{DD}$ during read, as shown by the shaded regions. The bit-cell 'holds' data when *SEL* is *GND*.

allow current to flow from SLL ($I_{READ,L}$) and from SLR ($I_{READ,R}$) to BL. $I_{READ,L}$ or $I_{READ,R}$ will be larger depending on the data stored in the bit-cell. If the selected CPSTT cell stores a '0', the current path from M7–M9 will be stronger than that from M3–M5. Hence, it is easier to charge up *SDB* than *SD* to V_{DD}. On the other hand, the current path from M3–M5 will be stronger than that from M7–M9 if the selected CPSTT cell stores a '1'. Then, it is easier to charge up *SD* than *SDB* to V_{DD}. At the end of the clock cycle,the voltages of *SDB* and *SD* will be V_{DD} and GND (GND and V_{DD}), respectively, if the selected CPSTT cell stores a '0' ('1'). The result is then latched into the D flip-flops at the end of the cycle.

6.2.2 Evaluation of bit-cells using complementary polarizer MTJ

The complementary polarizer MTJ (CPMTJ) structure was proposed in the previous section. The proposed structure avoids the source degeneration problem during write operations and enables self-referenced differential sensing for read operations. In this section, the CPMTJ based STT-MRAM (CPSTT) bit-cell is evaluated using the simulation framework described in Chapter 2. The layout for the CPSTT bit-cell is first presented alongside the layout for Standard STT-MRAM bit-Cell (SSC) so that the CPSTT bit-cells may be compared to SSCs at the same bit-cell layout area. The read and write performance of CPSTT bit-cells are then compared to those of SSCs.

Layout comparisons

The layouts for Standard STT-MRAM bit-Cell (SSC) and CPSTT shown in Fig. 6.10 and Fig. 6.11 respectively are drawn using λ based layout rules [69, 70]. As Fig. 6.11(d) shows, the area of the memory cell may be limited by the minimum metal pitch when the required ATx width is small. Thus, the cell area may remain constant when ATx width changes, such as for SSCs with ATx width below 200 nm. On the other hand, the fingered transistor layout may be used to reduce parasitics by

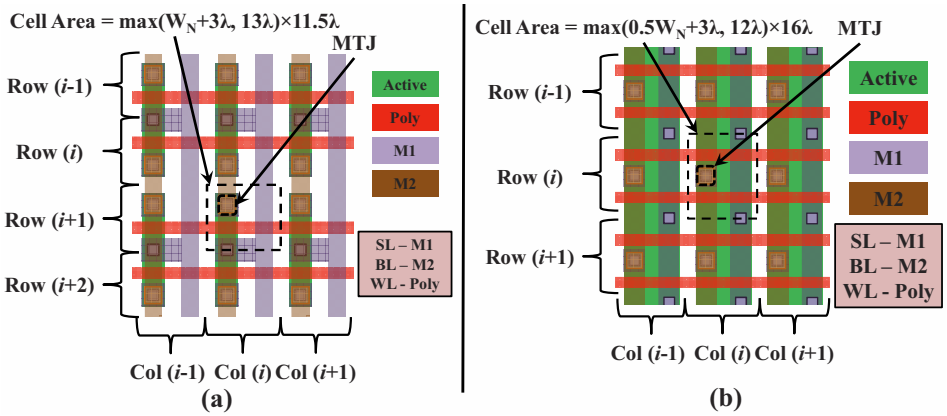


Fig. 6.10. Layout of Standard STT-MRAM bit-Cells (SSCs) (a) without and (b) with fingered ATx. SSC Layout without fingered ATx may be limited by the metal pitch as shown in (a). The layout in (b) is identical to that of 2T-1MTJ STT-MRAM bit-cells with shared WL.

implementing a single large transistor as several smaller transistors when the transistor is sufficiently large [1]. However, the layout area for fingered transistors may be limited by the metal pitch just like when ATx width is too small. Fig. 6.11(d) plots the memory cell area versus ATx width for CPSTT memory cells and SSCs. Because the minimum allowed ATx width is 120 nm in the CMOS technology used in this analysis, the fingered ATx layout for CPSTT may only be used for ATx width $\geq$240 nm. Consequently, the CPSTT cell area for ATx width between 180 nm and 239 nm is larger than that for ATx width of 240 nm. Hence, increasing ATx width to optimize CPSTT and SSC bit-cells may be done without cell area penalty within certain ranges of ATx width. The bit-cell area is fixed at 0.1152 μm^2 in the following comparisons between CPSTT and SSC.

Comparison of write performance

The parameters for bit-cell level simulations of SSCs and CPSTT are shown in Table 6.2. The complementary PLs in CPSTT need to be separated by an amount dependent on the layout rules. Hence, the FL in CPSTT is enlarged to allow it

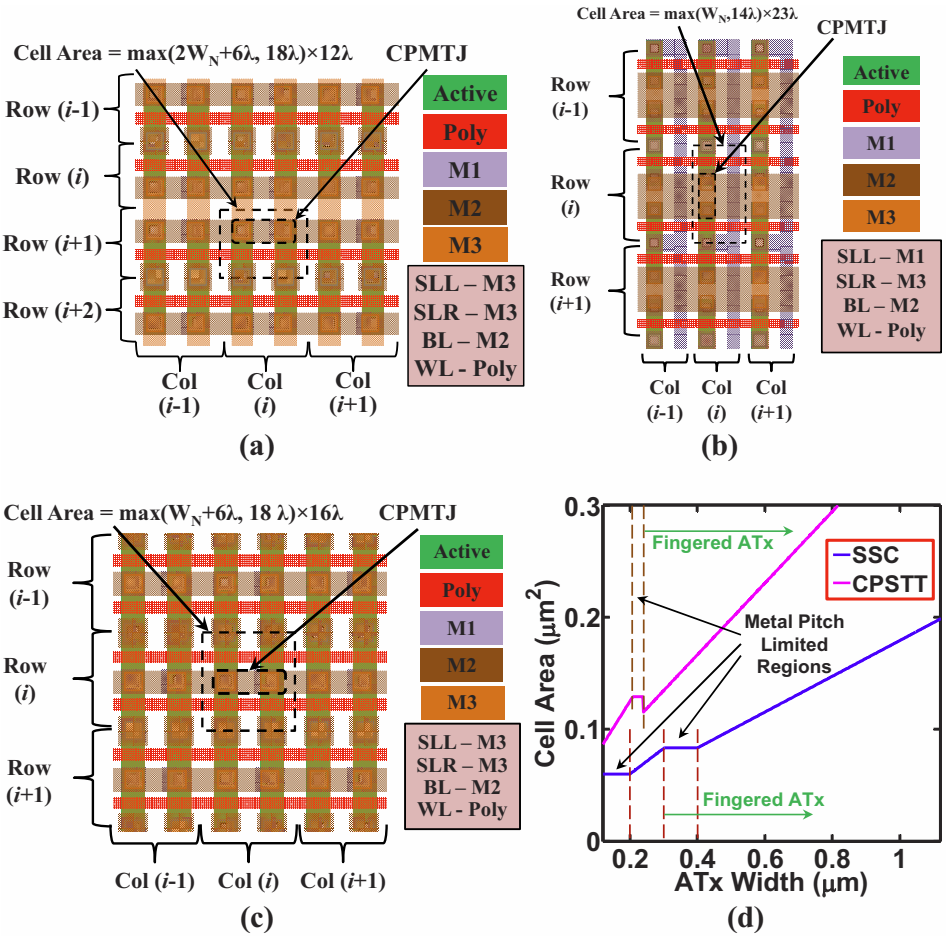


Fig. 6.11. Different layouts of the CPSTT bit-cell explored in this work are shown in (a) and (b). The fingered ATx layout in (c) is used when the ATx width is large. The comparison of CPSTT and SSC bit-cell areas at iso-ATx width is shown in (d). The metal pitch limited region for CPSTT corresponds to the layout in (b). The layouts for SSC are shown in Fig. 6.10.

to interface with both PLs. As a result, I_C('0') of CPSTT is larger than that in SSC. However, SSC requires bi-directional write current flow to program the bit-cells, whereas CPSTT always parallelizes the FL with a PL. Hence, I_C('1') of CPSTT can be lower than that of SSC, as shown in Table 6.2. Furthermore, the ATx's are never source degenerated during CPSTT write operations. Table 6.3 shows the V_{DD} required for CPSTT and SSC bit-cells to meet the required write margins (defined

Table 6.2.
Simulation Parameters for Bit-cell Comparisons

Retention Barrier Height	$56k_BT$
Write pulse width	2.0 ns
FL size (SSC)	10 nm × 10 nm × 1.5 nm
FL size (CPSTT)	10 nm × 22.5 nm × 1.5 nm
$TMR,\ RA_P$ at $V_{MTJ} = 0$ V	160%, 7.5 $\Omega - \mu$m^2 at $t_{MgO} = 1.15$ nm
Bit-cell Area (SSC and CPSTT)	0.1152 μm^2
ATx Width (SSC, CPSTT)	600 nm, 240 nm
t_{MgO}	1.0 nm
CMOS Technology	45 nm bulk CMOS
SSC: I_C('0'), I_C('1')	8 μA, 16 μA
CPSTT: I_C('0'), I_C('1')	13.5 μA, 13.5 μA

I_C was calculated from 1300 OOMMF monodomain simulations at $T = 300$ K

Table 6.3.
Iso-Write Margin V_{DD} and Average Write Power Per Bit

Write Margin	SSC	CPSTT	SV-CPSTT
0%	0.700 V, 11.92 μW	0.581 V, 11.59 μW	0.553 V, 10.19 μW
10%	0.738 V, 13.62 μW	0.618 V, 13.53 μW	0.588 V, 11.95 μW
20%	0.775 V, 15.39 μW	0.655 V, 15.52 μW	0.623 V,13.75 μW

as write margin = $\frac{I_{WRITE} - I_C}{I_C}$), and the corresponding average write power per bit. CPSTT write operations consume less power per bit than SSC at iso-write margin and iso- bit-cell area because of two reasons. Firstly, the access transistors are not source degenerated. Secondly, V_{DD} may be lowered to meet I_C requirements at the same bit-cell area for iso-performance.

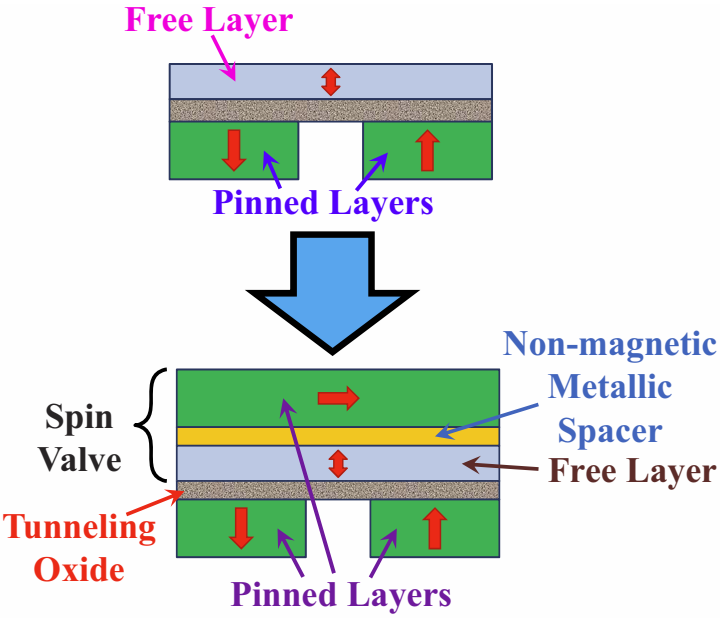


Fig. 6.12. The inclusion of a spin valve (SV) structure may reduce I_C of CPSTT.

However, I_C may still be too high due to the larger FL volume in CPSTT as compared to in SSC. Table 6.3 shows that as a result of the large I_C, CPSTT may still dissipate higher write energy per bit than SSC at large write margins. The I_C for switching FL at GHz speeds may also be prohibitively large. Several works have proposed reducing I_C in conventional MTJs by replacing the FL with a GMR based spin valve structure [71–73]. The FL in CPSTT may also be replaced with a spin valve (SV) as illustrated by Fig. 6.12, to reduce I_C. The modified CPSTT structure is denoted as SV-CPSTT. During write operations, current flows through the top PL through a non-magnetic metallic spacer (which may be Cu) before entering the FL. The current then tunnels across the tunneling oxide into one of the bottom PL just like in the basic CPSTT. Note the top PL is magnetized perpendicular to the easy directions of the other magnetic layers. The spin torque acting on FL due to the top PL provides a large initial torque that aids the switching of FL magnetization and hence, leads to reduced I_C [45, 71].

Table 6.4.
Iso-V_{READ} Comparison of Sensing Margins At $V_{DD} = 1.0V$

$V_{READ} = 0.3$ V	SSC	CPSTT
I_{REF}	9.57 μA	6.50 μA
$I_{READ,P}$	12.53 μA	12.17 μA
$I_{READ,AP}$	6.60 μA	6.50 μA
Margin	31%	87%
Avg. Read Energy / Bit	11.48 fJ	5.60 fJ

Table 6.5.
Iso-V_{READ} Comparison of Disturb Margins At $V_{DD} = 1.0V$

SSC	CPSTT
3.47 μA $= 0.217 \times I_C$	11.83 μA $= 0.493 \times I_C$

Table 6.3 also shows the V_{DD} and average write power per bit for SV-CPSTT under the same simulation conditions as the other memory cells. The additional torque from the orthogonal PL reduces I_C from 13.5 μA to 12.5 μA and also reduces the required V_{DD} to meet the same write margin at iso-cell area. Thus, the SV structure lowers the write power dissipated by CPSTT when large write margins are required. Results in Table 6.3 show that SV-CPSTT has 10%–14% lower write power than SSC at iso-write margin and iso- bit-cell area.

Comparison of read performance

Table 6.4 and 6.5 summarizes the read performance of SSC and CPSTT. Instead of implementing the sense amplifier in Fig. 6.8, the comparison was done using D.C. current sensing scheme for both SSC and CPSTT. $|V_{SL} - V_{BL}| = V_{READ} = 0.3V$ was assumed for SSC and $V_{SLL} - V_{BL} = V_{SLR} - V_{BL} = V_{READ} = 0.3V$ was assumed for CPSTT. The reference current for SSC was calculated as $I_{REF} = 0.5 \times$

($I_{READ,AP} + I_{READ,P}$). Since the self-referenced differential sensing scheme in CP-STT compares the two read currents through the bit-cell, I_{REF} for CPSTT is defined as the current flowing through the PL that is anti-parallel to the FL. Also, the sensing margin is defined as $\frac{I_{READ,P} - I_{REF}}{I_{REF}}$. Table 6.4 shows that the sensing margin is more than 2.8× that of SSC, and the average read energy per bit for CPSTT is 51.2% lower than in SSC. SSC has significantly higher read energy per bit because I_{REF} needs to be generated separately. Note that the sensing margin in SV-CPSTT is the same as that in CPSTT. Finally, Table 6.5 compares the disturb margins (defined as $|I_{READ} - I_C|$) in CPSTT and in SSC. In this comparison, the fact that the torque per read current in CPSTT is lower than in SSC is neglected. Hence, the disturb margin in CPSTT shown in Table 6.5 is the worst case disturb margin. Even so, the disturb margin of CPSTT is 3.4× that of SSC. However, the disturb margin for SV-CPSTT is 1.0 μA lower than that of CPSTT. Even so, the disturb margin for SV-CPSTT is 3.1× that of SSC. The disturb margins for CPSTT and SV-CPSTT are expected to be even better when the latch based sense amplifier in Fig. 6.8 is used for sensing CPSTT. A full transient simulation in SPICE with realistic parasitics was used to evaluate CPSTT read operation with the sense amplifier proposed in Fig. 6.8. SRAM bit-lines in 45 nm CMOS technology may have stray capacitances as high as 100 fF [74]. Hence, the stray capacitances on BL, SLL, and SLR are assumed to be 100 fF, and 1 pF on *SEL* in SPICE simulation of CPSTT using the periphery circuitry in Fig. 6.8. Transient SPICE simulation results show that read operations up to 1.5 GHz are possible, at read energy of 14 fJ/bit.

6.3 Cache Design using Complementary Polarizer MTJ

Processor performance is greatly improved by the use of caches [75]. The process of fetching data from off-chip may take hundreds to thousands of cycles and thus limits the performance of computing platforms. On-chip caches improve processor performance by storing copies of more frequently accessed memory locations closer

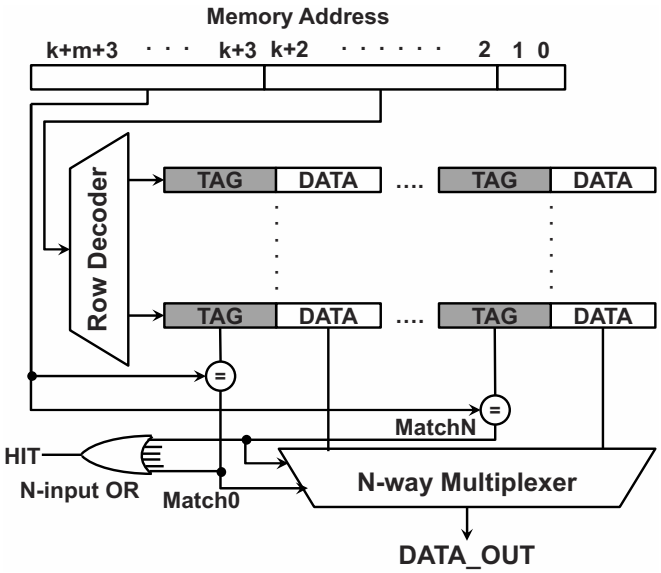


Fig. 6.13. Architecture of an N-way associative cache having $k+m+3$ bits wide address. There are N tag-data pairs per row of cache and 2^k number of rows. During read, the m most significant bits of the address are checked against the tag bits in the tag array to determine whether the cache contains a copy of data in stored memory. A cache *hit* (*miss*) occurs if data is (not) in cache.

to the processor cores. This section shows how the proposed CPSTT can be used in the design of on-chip caches.

6.3.1 The tag array

Since cache is a small chunk of very fast memory, the processor needs to map memory addresses in cache to memory addresses in main memory. In associative caches, the cache location corresponding data memory address is stored in the tag array. The memory address stored in the tag array, together with the tag address where the tag is stored, forms the address of the memory location in main memory (Fig. 6.13). When the processor accesses a memory location, the memory controller checks in cache first to see if the data corresponding to that memory location is already loaded in cache. If the data is not already in cache, it will then check in the

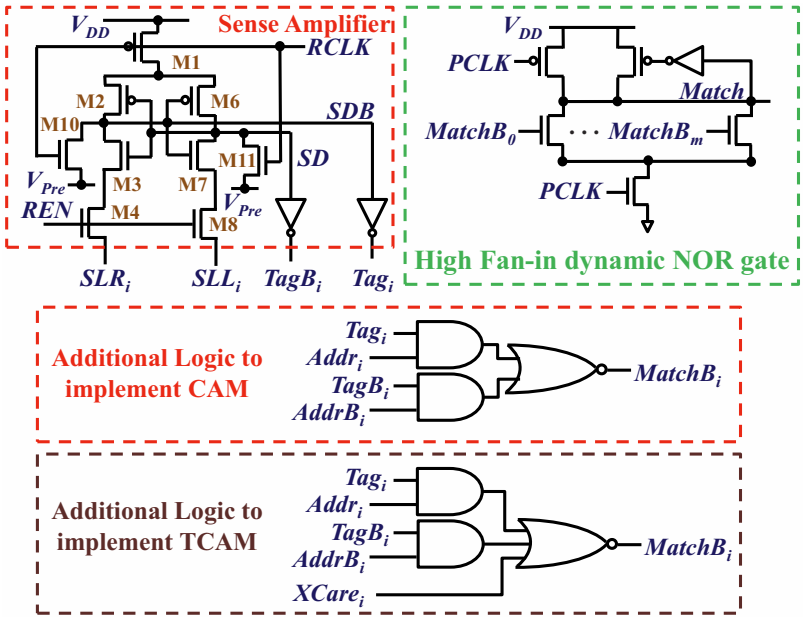


Fig. 6.14. Additional logic is added to the sense amplifier from Fig. 6.8 to implement CPSTT based content addressable memory (CAM). The sense amplifier of the i-th column (or bit) in the row is shown here, with *Data* and *DataB* renamed to Tag_i and $TagB_i$, respectively. Every bit in the tag in Fig. 6.13 is compared to the corresponding bit in the m most significant address bits using the additional logic shown for CAM and/or ternary CAM (TCAM). The result of each bit comparison goes to a high fan-in dynamic NOR gate shown. The output of the NOR gate goes into the input of the OR gate shown in Fig. 6.13 to determine whether there is a cache hit.

next level of memory hierarchy and so on until it finds the data [75]. Once the data is found, the memory controller copies it into cache and the processor can continue program execution. Thus, the tag array stores address bits that need to be compared during every memory access and the system only needs to know if the contents of the tag match part of the address to main memory. Such a memory structure is called *content addressable memory* or CAM [1]. Note that bit-comparison in CAM is done for all the bits in a row. On the other hand, a ternary CAM, or TCAM, is a special kind of CAM in which bit-comparison on some bits in the row can be ignored. A "don't care" signal tells the TCAM which bit-comparisons may be ignored. Thus,

the difference in the array storing data and the array storing tags in the cache is the additional logic required to compare stored tag bits to the address bits as shown in Fig. 6.14.

When checking if every tag bit matches the corresponding address bit, a signal to indicate a match is generated for every bit-comparison. The signals are AND-ed together to determine if the tag matches the address. High fan-in AND logic gates tend to be very slow and can significantly degrade cache performance. Alternatively, the same bit-comparison can be done by checking if any tag bit does not match the corresponding address bit. Any mismatch indicates that tag and address are not the same. The signal for every comparison can be NOR-ed together to determine if the tag matches the address. A fast circuit implementation for high fan-in NOR gates uses dynamic style logic [1] as shown in Fig. 6.14.

In a CPSTT based CAM, the memory cells do not have to be modified. Since each bit-position corresponds to a column in memory, additional logic may be integrated into the sense amplifier for every column to compare the stored tag bit with the corresponding address bit [see Fig. 6.13 and Fig. 6.14]. The sense amplifier in Fig. 6.8 is modified to enable CAM and TCAM capabilities as shown in Fig. 6.14, where Tag_i is the i-th bit stored in a row in the tag array. Due to the differential nature of the sense amplifier, complementary signals (Tag_i and $TagB_i$) are available. Comparison with the i-th address bit can be done using the logic shown in Fig. 6.14. If Tag_i does not match the corresponding address bit ($Addr_i$), both AND gates output logic '0'. However, one of the AND gates will output logic '1' if Tag_i matches $Addr_i$. A high fan-in dynamic NOR gate (Fig. 6.14) checks if any of the m tag bits do not match the corresponding address bit. *Match* is preset to logic '1' when *PCLK* is '0' during the preset phase. When *PCLK* goes to '1' in the evaluation phase, *Match* will be pulled down to logic '0' if any of the $MatchB_i$ signals is '1'. This *Match* signal goes to the N-input OR gate in Fig. 6.13, which tells the cache controller whether data corresponding to the memory address is found in cache (a cache *hit* if it is, and cache *miss* if not).

In a CPSTT based TCAM, an additional input, $XCare_i$, controls whether the i-th tag bit comparison with the corresponding address bit is ignored. TCAM can be enabled by adding one more input to the NOR gate used for CAM as shown in Fig. 6.14. If $XCare_i$ is high, $MatchB_i$ will be '0' regardless of Tag_i and $Addr_i$. Hence, $Match$ becomes independent of the i-th bit comparison if $XCare_i$ is '1'.

The CAM and TCAM were implemented using CPSTT and validated in SPICE. Simulation parameters for the CPSTT cell used in the CAM and TCAM are the same as those presented in the previous sections. The additional logic gates shown in Fig. 6.14 are implemented as single-stage CMOS logic gates instead of multi-stage gates. In the CMOS technology used for this work, the single-stage gate delay is about the same as the delay of a two-input NAND gate. Transient SPICE simulation results show that CAM/TCAM operations at frequencies up to 1.5GHz are possible. Since both CAM/TCAM and RAM read operations may be clocked at 1.5GHz, CPSTT based on-chip L1 caches may be implemented with latencies comparable to SRAM based on-chip L1 caches.

6.3.2 Column-selection

When the data being accessed is located in cache, the cache read operation proceeds as follows (Fig. 6.13). The row of bit-cells corresponding to the memory address is accessed. A tag search is performed in corresponding row of the tag array. Since the data of the corresponding memory location is already in cache, the tag search returns a hit, and the MATCH corresponding to the tag location is asserted. The N-way analog multiplexer now connects the write drivers and the sense amplifiers to the source lines of the corresponding columns. The cache access scheme just described is called the *sequential tag-data access* (Fig. 6.15), where data sensing is done only after the tag search returns a hit [75]. The alternative access scheme is the *parallel tag-data access* (Fig. 6.15) where data sensing on all the memory cells in the row are done in parallel with the tag search [76]. In order to reduce the number of sense

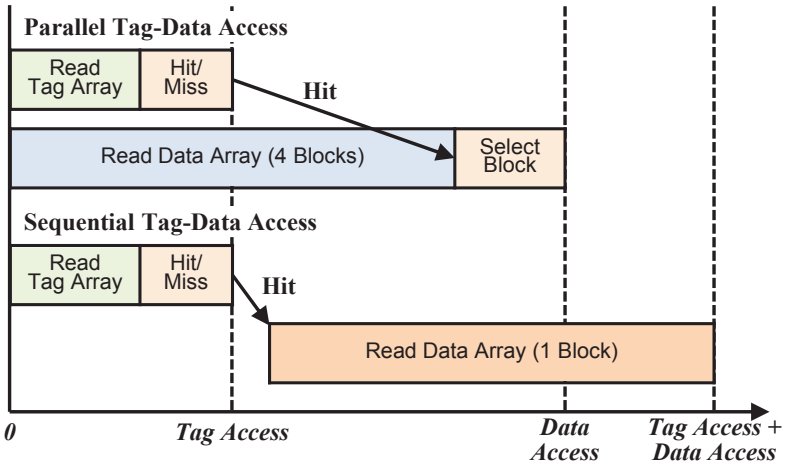


Fig. 6.15. Timing of (top) parallel and (bottom) sequential tag-data access.

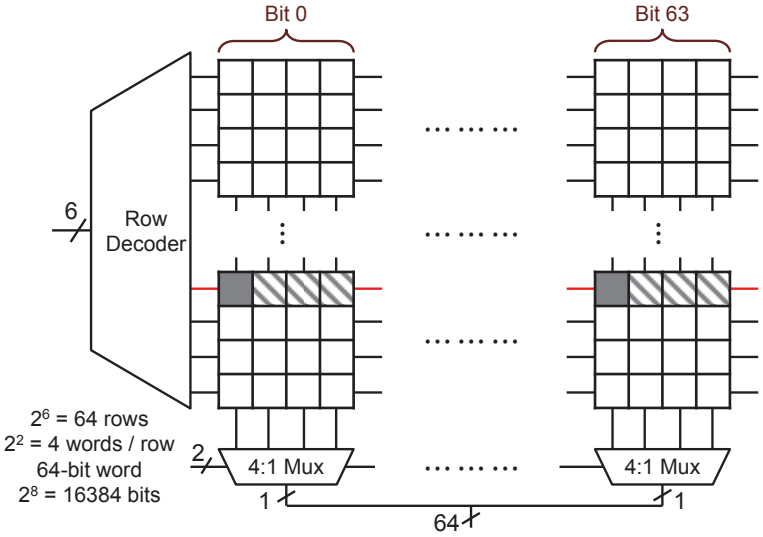


Fig. 6.16. Bit-interleaving reduces the multiplexer wiring as shown in this illustration using a 16kb (kb = kilobit) array storing 64 bit words with 4-way associativity. The n-th bit of each word is stored in four adjacent columns to reduce the wiring from the columns to the 4:1 multiplexers. When a word is being read out (solid shaded square), the word line of the selected row (red line) is turned on and the select signal to the multiplexers determine which of the four words stored in the row is read out.

Table 6.6.
Processor Configuration for System Simulation

Processor Core	Alpha 21264 pipeline, out-of-order, issue width-4
Functional Units	Integer - 8 ALUs, 4 multipliers Floating Point - 2 ALUs, 2 multipliers
L1 Data Cache	32-kilobytes, direct mapped, 32-byte line size
L1 Instruction Cache	32-kilobytes, direct mapped, 32-byte line size
L2 Unified Cache	2MB, 4-way associativity, 64-byte line size

amplifiers required for SSC and CPSTT based cache arrays, the sequential tag-data access scheme is used. Furthermore, the wiring from the read and write peripheral circuits is reduced using bit-interleaving [76–78] as illustrated in Fig. 6.16.

6.3.3 System-level evaluation of CPSTT based on-chip cache

The overall energy consumption, area, and performance of CPSTT based caches are compared to an SSC based cache using a modified version of the CACTI 6.5 cache modeling tool [77] and the SimpleScalar architectural simulator [79] for a wide range of SPEC2K6 benchmarks. The processor configuration used in our analysis is shown in Table 6.6. In this work, SSC, CPSTT, and SV-CPSTT bit-cells are implemented in both the tag and the data arrays of L2 cache. The L2 cache access is assumed to be sequential in which the tag is compared first and the data array is accessed only for hits as explained in Section 6.3.1. For an SSC-based tag array, the tag data has to be read out first and compared. The data array is then accessed if there is a hit. On the other hand, CPSTT based caches can read the tag data and perform comparisons in one cycle. The data array is then accessed if there is a hit. Therefore, the assumed read latency of the SSC cache is twice that of CPSTT cache.

For fair comparison, cache arrays based on SSC, CPSTT and SV-CPSTT are compared at iso-area, write margin and capacity (2 MB, MB = Mega Byte). Bit-cell

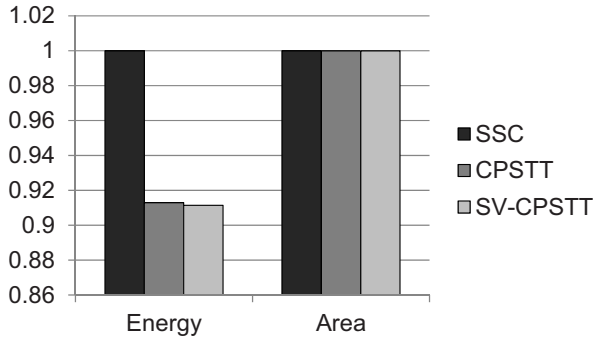


Fig. 6.17. Energy consumption and area comparison of 2 MB (MB = Mega Byte) L2 cache based on SSC, CPSTT, and SV-CPSTT. The results are based on the bit-cell level results for 20% write margin in Table 6.2 to 6.3.

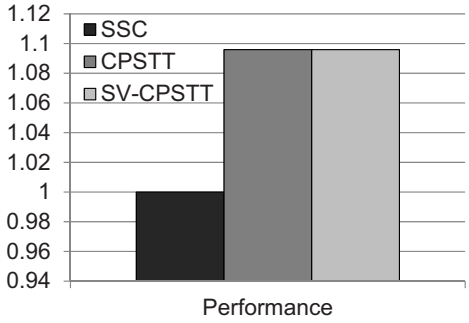


Fig. 6.18. Performance comparison of 2 MB (MB = Mega Byte) L2 cache based on SSC, CPSTT, and SV-CPSTT, based on bit-cell level results for 20% write margin in Table 6.2 to 6.3.

level parameters used to obtain the results are tabulated in Table 6.2 and 6.3. Fig. 6.17 shows that the total energy consumption of a CPSTT based cache is $\sim 9\%$ lower than that of an SSC based cache even though the write power per bit-cell is substantially lower. The modest energy improvement in CPSTT based caches stems from three factors: 1) write operations do not occur as often as read operations, 2) source-lines of unselected bit-cells need to be charged to avoid disturbing them when writing into the selected bit-cells, and 3) energy consumption is dominated by charging of the word-lines and bit-lines. Furthermore, Fig. 6.18 shows that CPSTT based caches achieve $> 9\%$ higher Instructions Per Cycle (IPC) than SSC based caches due to much lower

cache read latencies. As shown in [31], SSCs may require multi-cycle read operations to mitigate sensing errors. Thus, cache performance, which is very sensitive to read latency, is much better in CPSTT based caches than in SSC based caches.

6.4 Summary

In the earlier chapters, it was shown that the key design issues hindering standard STT-MRAM–shared read and write current paths, source degeneration of the access transistor during write operations, and single-ended sensing operations–arise due to the two-terminal nature of the storage device (the magnetic tunnel junction, MTJ). This chapter described the multi-terminal MTJ structures proposed in the literature and proposed a novel complementary polarizer MTJ (CPMTJ) structure that overcomes all the aforementioned design issues in STT-MRAM. The evaluation of the CPMTJ based STT-MRAM (CPSTT) bit-cell presented in this chapter showed that the average write energy in the CPSTT bit-cell may be increased due to an enlarged free layer. However, a spin-valve structure added to the CPMTJ (which is called the SV-CPSTT) may achieve 10% savings in average write energy. Since write operations may occur infrequently, a system-level evaluation was performed. The design of CPSTT caches was discussed first before the evaluation results were presented. Simulation results show that when the write margins are fixed and the array area and capacity are kept the same, CPSTT and SV-CPSTT based caches can achieve 9% improvement in performance and $> 8\%$ savings in energy consumption as compared to cache based on the standard STT-MRAM bit-cell.

7. ON-CHIP APPLICATIONS OF STT-MRAM

The design of robust STT-MRAM based on-chip caches has been discussed in the preceding chapters. Results from Chapter 6 show that STT-MRAM based on-chip cache may provide system-level benefits in terms of improvements in energy consumption and in performance. However, the most significant system level implication of STT-MRAM is that it allows new functions to be embedded within the on-chip caches with little to no area overhead, and no degradation in cache performance. This chapter explores two on-chip cache applications that may be significantly improved by STT-MRAM–in the area of on-chip hardware security (Section 7.1) and another in the area of application acceleration (Section 7.2).

7.1 STT-MRAM Based Random Number Generators

Random numbers are useful in security applications such as for cryptographic key generation as well as other applications such as Monte Carlo simulations. A spin dice was proposed as a 1-bit true random number generator (TRNG) implemented using standard STT-MRAM [80], shown in Fig. 7.1 (which is called cSD), and an m-bit random number may be generated by concatenating m spin dies. The operation of cSD requires three sequential steps as illustrated in Fig. 7.2: 1) *initializing* or *resetting* the cSD to a known state; 2) stochastic programming of the cSD by current-driven STT (also called *rolling* the dice); and 3) *sensing* the final state of the cSD. However, several design issues limit the efficacy of cSD. Steps 1 and 2 are required to randomize the state of the cSD. The final state of cSD is sensed by passing a current through the magnetic tunnel junction (MTJ) in the cSD. Under thermal fluctuations and process variations, the current flowing through the MTJ during sensing may bias the final state of the cSD in a similar way that STT-MRAM is affected by the read

disturbance problem, which was discussed earlier in Section 3.1.2 and in [21, 38]. Hence, the randomness cSD is degraded because the current paths for programming and sensing are shared. Increasing the activation energy (E_A) of the MTJ increases the current required to flip the MTJ state during sensing operations and mitigates the sensing bias in cSD. However, doing so increases the critical switching current (I_C) needed to program the MTJ and hence, increase the power consumed by the cSD.

7.1.1 CPSTT based TRNG

The complementary polarizer STT MTJ (CPMTJ) structure discussed in Chapter 6 may be used to implement on-chip spin dice (CPSD, shown in Fig. 6.6 and repeated here in Fig. 7.3). The CPSD structure overcomes the design issue in cSD by enabling self-referenced differential sensing of the CPSD state. First, consider the operation of the CPSD. Fig. 7.4 shows that the state of the CPSD is reset by passing a current from the bit-line (BL) to the left source-line (SLL). Rolling of the CPSD

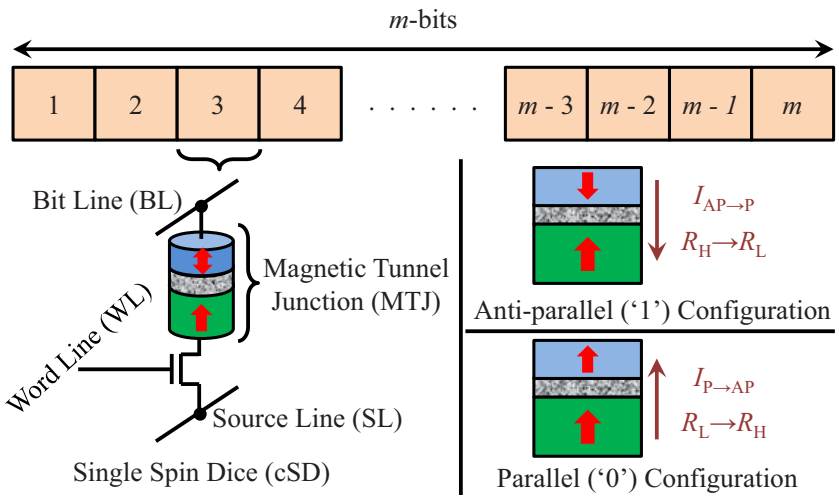


Fig. 7.1. Schematic diagram of an m-bit random number generator implemented using STT-MRAM based spin dice. The directions of current flow through the MTJ to program it are shown on the right.

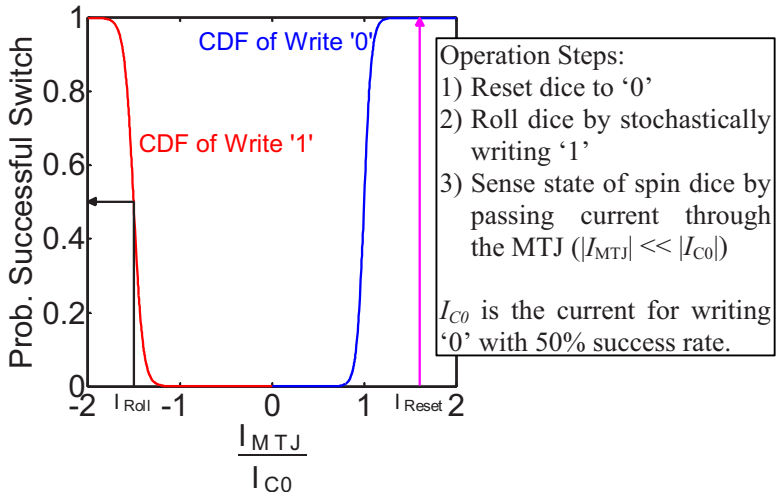


Fig. 7.2. Illustration of spin dice operation using an example CDF of MTJ switching characteristics. The stochastic nature of spin-transfer torque is exploited to generate '1' with 50% probability.

is done by passing current from BL to the right source-line (SLR). The state of the CPSD is sensed by first biasing the CPSD as shown in Fig. 7.5 and then comparing the current flowing from BL to SLL and BL to SLR. Consider when the FL is more closely aligned with SLL than SLR (Fig. 7.6). The current flowing out of SLL will be larger than that through SLR. In the monodomain limit, the net torque acting on the FL due to the two currents acts to align the magnetization of FL in the direction of the PL that is connected to SLL. By similar arguments, the FL becomes aligned with PL connected to SLR if the current flowing from BL to SLR is larger than that flowing from BL to SLL during sensing. Hence, there is a positive feedback loop that stabilizes the FL magnetization during sensing and preserves the randomness of the CPSD. Note that I_C for the CPMTJ may be larger than that of the conventional MTJ because of a larger FL (which was explained in Chapter 6). Since the primary contribution to power consumed by a spin dice is the power consumed to reset and to stochastically program the spin dice, the larger I_C of CPMTJ may result in higher power consumption of CPSD compared to cSD. Note that E_A is required in cSD to preserve randomness. Since the sensing operation in CPSD is stable, the energy bar-

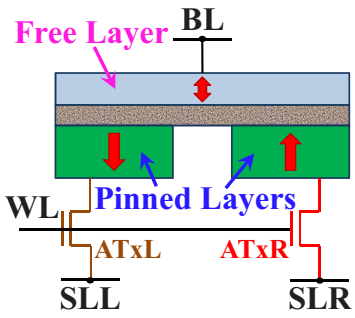


Fig. 7.3. The structure of the complementary polarizer STT-MRAM bit-cell which may be used as a spin dice.

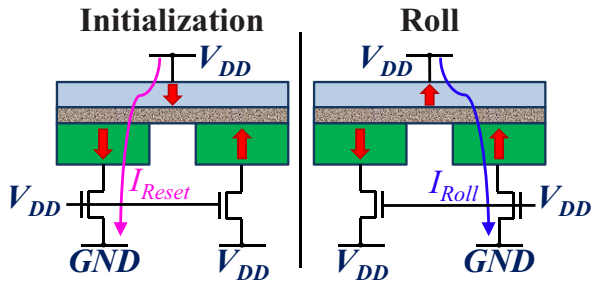


Fig. 7.4. Direction of current flow in the CPSD for (left) the *reset* or *initialization* operation, and (right) the *roll* operation.

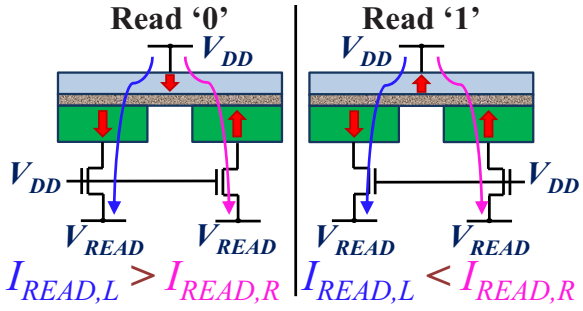


Fig. 7.5. Voltage bias and current flow through the CPSD during sensing operations.

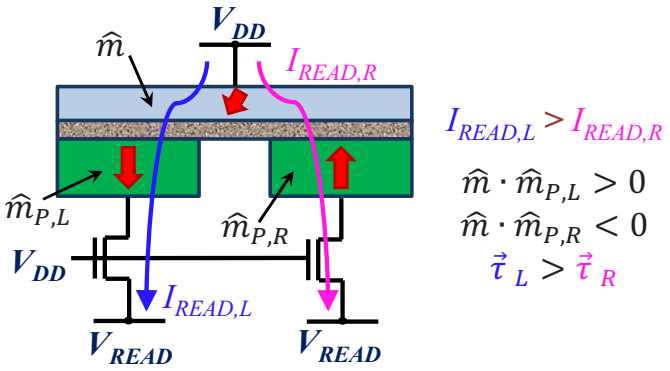


Fig. 7.6. The net torque due to the currents flowing through the left and right PL's, $\vec{\tau}_L$ and $\vec{\tau}_R$, respectively, tries to align the FL magnetization ($\hat{m}$) with the closest PL magnetization ($\hat{m}_{P,L}$ here).

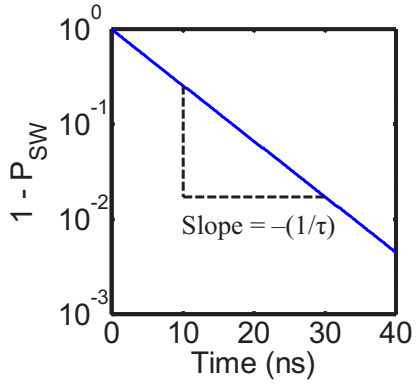


Fig. 7.7. The randomness of the CPSD depends on the frequency of operation as shown by the switching probability versus time, $P_{SW} \propto e^{-\frac{t}{\tau}}$.

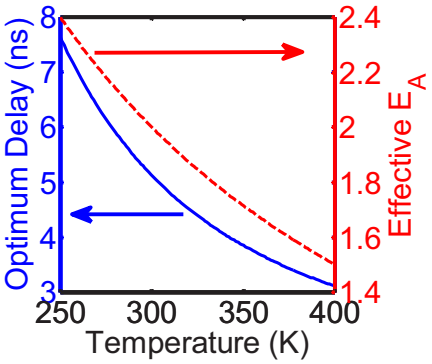


Fig. 7.8. The optimum sensing delay and effective E_A of CPSD versus the operating temperature. The randomness of the CPSD may hence be degraded by fluctuations in temperature and process variations.

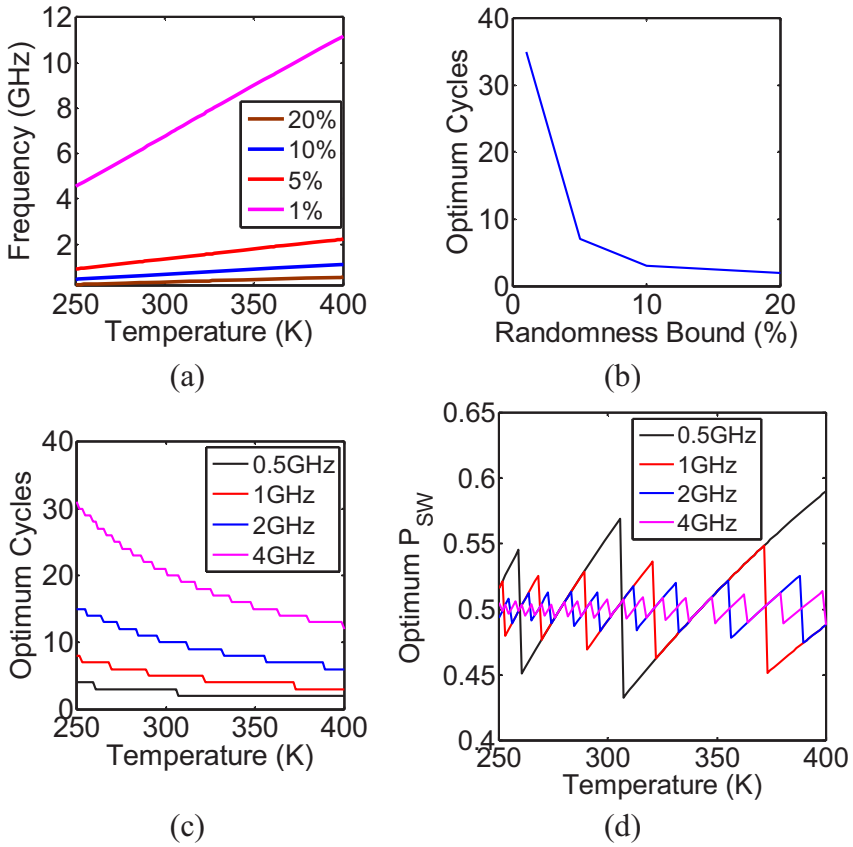


Fig. 7.9. The robustness of CPSD against temperature fluctuations may be enhanced by tuning the operating frequency. (a) plots the dependence of operating frequency on temperature for different levels of randomness (i.e., P_{SW} is within XX% of 0.5). (b) shows the optimum number of cycles between CPSD sensing events depends only on the level of randomness and not on the operating temperature. However, high operating frequencies may be difficult to achieve. If operating frequencies are fixed, the number of cycles between CPSD sensing events can be tuned to optimize the CPSD randomness with varying temperature as shown in (c) and (d). The achievable levels of randomness at different temperatures for different CPSD operating frequencies are shown in (d). Since the CPSD footprint is small, sequential access of an array of CPSD may be used to improve the throughput of random number generation. Each row of m CPSD cells generates a random m-bit word. n rows of CPSD cells, accessed sequentially automatically imposes a delay between consecutive access to the same row of CPSD cells. Ideally, n should match the optimum number of cycles between consecutive accesses to the same row of CPSD cells.

rier required may be much lower than in cSD and may be lowered to reduce power consumption. Furthermore, E_A determines the frequency of random switching events in the MTJ due to thermal fluctuations (lower E_A increases frequency of random switching events). If the frequency of random switching events is sufficiently high, the CPSD state may be randomized using thermal fluctuations instead, eliminating the need for reset and programming operations. Hence, CPSD may be an energy efficient on-chip true random number generator.

7.1.2 Evaluation of CPSTT based TRNG

The characteristic switching time [81] of a CPSD may be calculated as

$$\tau = t_0 e^{\frac{E_A}{k_B T}} \tag{7.1}$$

where k_B is the Boltzmann constant and T is the temperature of operation. The switching probability (P_{SW}) of the CPSD is plotted in Fig. 7.7. The randomness of the CPSD is optimized with $P_{SW} = 0.5$. However, the effective E_A and the randomness of the CPSD changes with temperature. Maximizing the CPSD randomness limits the throughput of random number generation (Fig. 7.8). Since the CPSD footprint is small, an array of CPSD cells accessed sequentially allows CPSD randomness to be maximized without degrading the throughput. However, the operating frequency of such an array needs to be very high for increasing levels of randomness, as shown in Fig. 7.9(a). Fig. 7.9(b) shows that the optimum number of cycles between consecutive accesses to the same CPSD does not depend on temperature. If the operating frequency is fixed, as Fig. 7.9(c)-(d) shows, additional peripheral circuits and a hashing function are needed to maximize randomness.

7.2 Accelerating Applications using STT-MRAM

Many applications use data stored in the form of look-up tables. For example, math libraries are commonly used for the evaluation of complex math functions. Since

these libraries are usually stored off-chip, a significant amount of memory accesses take place when complex math functions are first called or when there are cache misses. Consider for example the first call to a complex math function during the execution of a computer program. A mandatory cache miss occurs and the processor needs to fetch the required look-up table data from the off-chip memory, which takes hundreds of clock cycles. Furthermore, the data already in cache may need to be moved out to accommodate the look-up table. As a result, the evaluation of such math functions may incur significant number of clock cycles before completion. One method to accelerate the evaluations of complex math functions is to store the look-up tables in on-chip read-only memory (ROM). However, the size of these look-up tables depends on the required accuracy of the math function evaluation larger look-up tables are needed for more accurate math function evaluation results. Hence, large ROMs may be required to accelerate the evaluation with the desired accuracy. The area required for these large standalone ROMs makes it impractical for on-chip implementation.

Since the size of on-chip cache (random access memory or RAM) in modern microprocessors may be as large as 8MB (MB = Mega-Byte), a method for embedding ROMs in on-chip caches (in other words, the ROM and cache area are shared) with little area overhead and performance penalty is desirable. This enables a practical implementation of ROMs for accelerating the evaluation of math functions. A method for embedding ROMs in SRAM based on-chip cache was presented in [82]. The authors report $\sim$ 30% improvement in evaluation latency for double-precision elementary math functions using ROM-embedded SRAM (R-SRAM) over conventional evaluation techniques. The RAM capacity of the R-SRAM is not impacted by the embedded ROM. In fact, the ROM capacity can be as large as the RAM capacity. Furthermore, the total area of R-SRAM is much smaller than the total area of the same implementation using separate iso-capacity RAM and ROM. However, additional buffer storage is needed to allow proper operation of R-SRAM as will be

described later. Hence, R-SRAM may suffer from high memory traffic that limits the improvement in evaluation latency of complex math functions.

R-SRAM may be viewed as a special type of resettable RAM. When ROM data at a corresponding memory location is needed, the RAM data at the corresponding memory location is overwritten with ROM data [82]. Hence, the RAM data is first copied to a buffer prior to the reset. The reset operation is then performed in one clock cycle and the ROM data is read out in the following cycle. Finally, RAM data is copied back into the memory location from the buffer. Hence, the latency of function evaluation in R-SRAM may still be high when RAM data and ROM data at the same memory address are frequently accessed. Consequently, the improvement in evaluation latency of math function using R-SRAM may be significantly lower than reported in [82].

Recently, spin-transfer torque MRAM (STT-MRAM) has emerged as the leading technology candidate for non-volatile on-chip cache memory [12]. STT-MRAM based cache may offer as much as $3\times$ higher capacity as SRAM based cache at iso-array area [76]. Furthermore, a methodology for embedding of ROM in STT-MRAM was proposed in [83]. Due to the non-volatility of STT-MRAM, ROM-embedded STT-MRAM (R-MRAM) behaves as a dual mode (RAM mode and ROM mode) memory system in contrast to R-SRAM. When ROM data is needed in R-MRAM, the RAM data is not overwritten, unlike in R-SRAM. Hence, in R-MRAM, there is no need to move RAM data to buffer storage when switching from RAM mode to ROM mode, and no need to restore RAM data from buffer storage when switching from ROM mode to RAM mode. The memory traffic from switching modes in R-MRAM is significantly lesser than in R-SRAM and hence, a dramatic improvement in evaluation latency of complex math functions may be achieved. Furthermore, R-MRAM may achieve much more accurate evaluation of complex math functions compared to R-SRAM because of the higher capacity at iso-array area.

The following sections propose Standard STT-MRAM bit-Cell (SSC) based and complementary polarizer STT-MRAM (CPSTT) based caches that can operate in

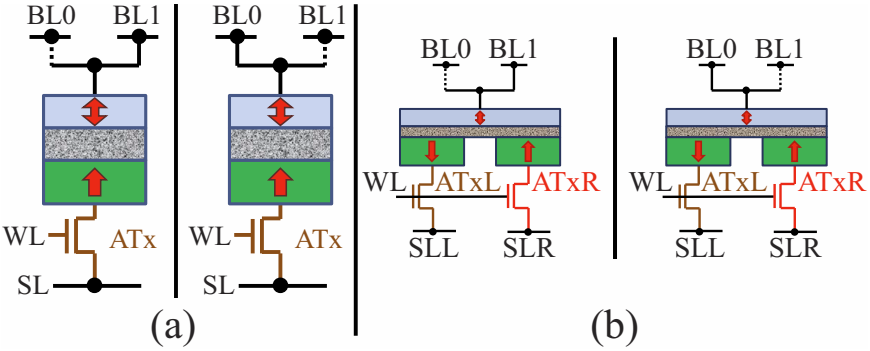


Fig. 7.10. Selective connection of (a) SSC and (b) CPSTT bit-cells to BL0 or BL1 allows ROM data to be programmed. Two bit-lines (BL0 and BL1) are needed but there is no area overhead when the ATx width is sufficiently large.

RAM mode or in ROM mode, which are called R-MRAM and R-CPSTT, respectively. Every bit-cell in R-MRAM and R-CPSTT is a single-level cell that stores both RAM and ROM data, which do not have to be the same. The MTJ structure in the bit-cell is used to store RAM data, whereas, as will be shown later, ROM data is stored as the selective connection of the bit-cell to one of two bit-lines. Data may be written to or read from any memory address during RAM mode of operation. In ROM mode of operation, only data that is programmed into the structure during design time may be read from any memory address. The proposed bit-cell designs do not compromise the density benefits of spin-based memories as will be shown later. The following section describes the R-MRAM and R-CPSTT in detail.

7.2.1 Embedding read-only memory in STT-MRAM

The key insight used to enable R-MRAM and R-CPSTT is the fact that an additional bit-line (BL) may be added to the cache arrays without bit-cell area penalty if the access transistor (ATx) is sufficiently large. ROM data may then be programmed as the selective connection of the bit-cell to one of the two available BL's (BL0 and BL1 in Fig. 7.10) during design time. During ROM mode operation, data is sensed

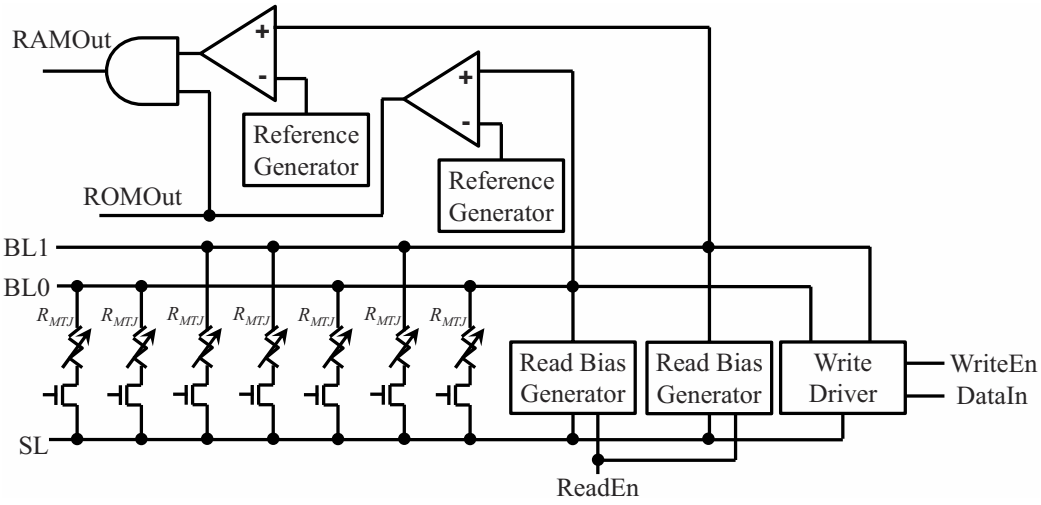


Fig. 7.11. Structure of the R-MRAM proposed in [83]. Every bit-cell may be programmed with RAM data. In addition, the physical connection of the bit-cell to BL0 or BL1 stores the ROM data. Bit-cells connected to BL0 store ROM data '0' whereas those connected to BL1 store ROM data '1'.

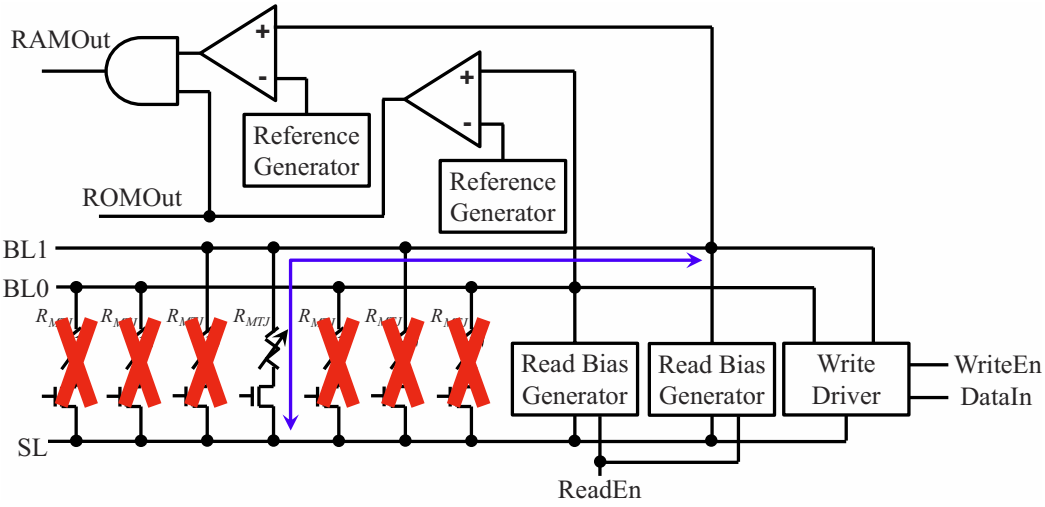


Fig. 7.12. Current flow in a selected bit-cell connected to BL1.

by determining whether the bit-cell is connected to BL0 or to BL1. On the other hand, BL0 and BL1 are electrically connected during RAM mode of operation. Note that ROM access and RAM access cannot occur simultaneously.

One design of ROM-embedded STT-MRAM was proposed in [83]. Fig. 7.11a shows a column of the R-MRAM array proposed in [83]. BL0, BL1 and SL are shared along the column of the array whereas WL is shared along a row. The R-MRAM array requires two sense amplifiers because BL0 and BL1 are not physically connected. Bit-cells that are connected to BL0 are programmed to store ROM data value of '0' whereas those connected to BL1 are programmed to store ROM data value of '1'. The WL is turned on to select a row of cells and current may flow through only one bit-cell in the column. During RAM write operations, the write driver ensures that both BL0 and BL1 are at the same voltage. The relative voltages of SL and the bit-lines depend on DataIn. During RAM read operations, SL is discharged to GND and the read bias generators act as a current source that drives current into BL0 and BL1. The sense amplifiers compare the voltage on the BL0 and BL1 to a common reference voltage, which is lower than V_{DD}. Note that the reference voltage depends on whether the ROM or RAM data is required. If the voltage on the BL is higher than the reference voltage, the sense amplifier outputs a '1', and '0' otherwise.

In the scenario shown in Fig. 7.12, the unselected cells in the column are marked with an 'X' and the selected bit-cell is connected to BL1. The output of the sense amplifier connected to BL1 depends on the resistance of the selected bit-cell. Since BL0 is a high impedance node, the current from the read bias generator charges BL0 to a voltage close to V_{DD}. Hence, the sense amplifier connected to BL0 will output a '1'. For a ROM read operation, the output of the sense amplifier connected to BL0 gives the result and is sent to the array output (ROMOut in Fig. 7.12). For a RAM read operation, the result of the read operation must be determined by the resistance of the selected bit-cell. The sense amplifier connected to the BL1 in Fig. 7.12 gives the correct result for the RAM read operation. However, if the selected bit-cell was connected to BL0 instead of BL1, the correct result of the RAM read operation is given

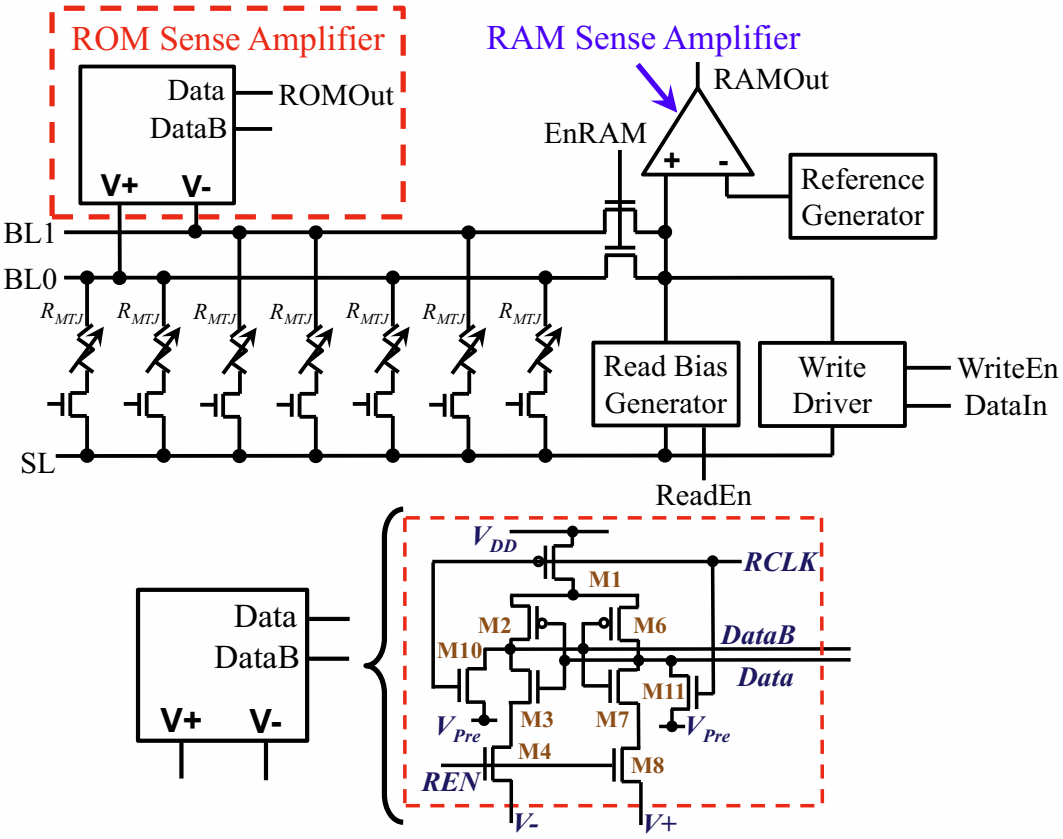


Fig. 7.13. The improved ROM-embedded MRAM proposed in this dissertation uses pass gates to electrically connect BL0 and BL1 during RAM mode operation so only one sense amplifier is needed for RAM mode read operations. ROM mode read operations use a latch to determine which bit-line is the high impedance node.

by the sense amplifier connected to BL0. Note that if a BL is a high impedance node, the sense amplifier connected to it will output '1' during read operations. During RAM read operation, one of the two sense amplifier will output '1' because the BL connected to it is the high impedance node. The output of the other sense amplifier depends on the resistance of the selected bit-cell. Hence, the result of the RAM read operation is obtained by AND-ing the outputs of both sense amplifiers (RAMOut in Fig. 7.12).

In the aforementioned design, both sense amplifiers need to be designed to reduce sensing failures during RAM mode of operation because the result of the RAM read

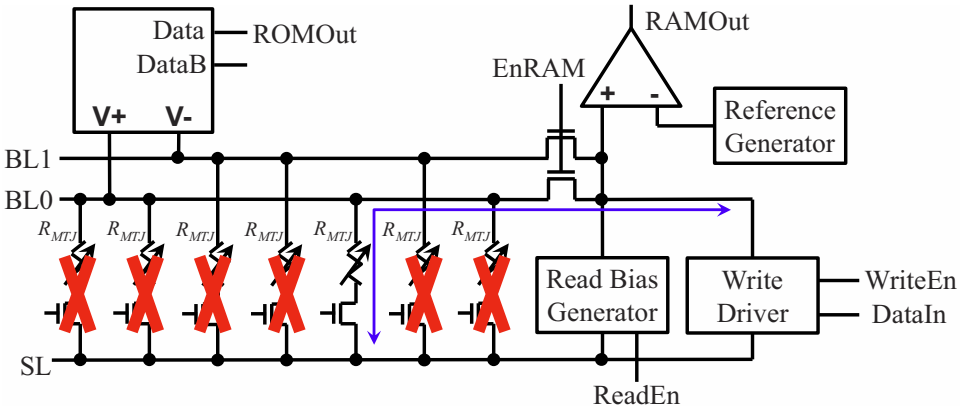


Fig. 7.14. Current flow in a selected bit-cell connected to BL0 during RAM mode operation.

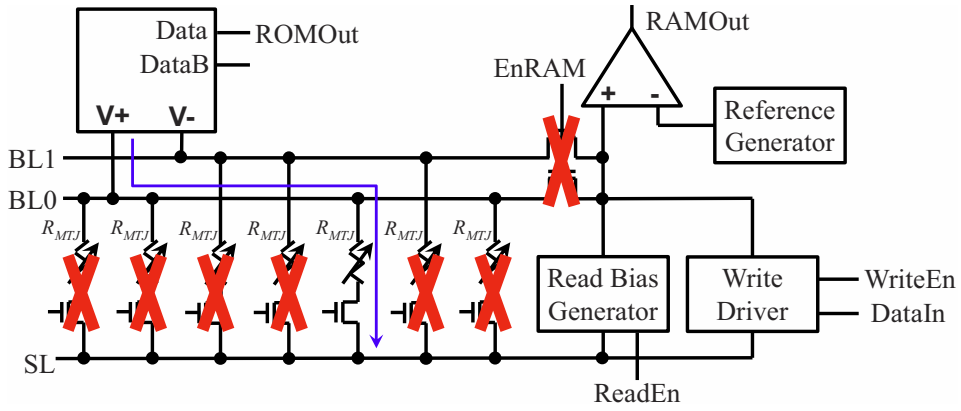


Fig. 7.15. Current flow in a selected bit-cell connected to BL0 during ROM mode operation.

operation can come from either of them. Thus, the area overhead from the sense amplifiers may be significant. Furthermore, the ROM mode read operation may be limited by the sensing speed of the sense amplifiers, which must meet RAM mode read operation requirements. To overcome these issues, we propose modifications to the peripheral circuitry as shown in Fig. 7.13. Two sense amplifiers are still needed but one is used exclusively for RAM mode read operations and the other is used exclusively for ROM mode read operations.

Consider the operation of the array when the selected bit-cell in the column is connected to BL0 as shown in Fig. 7.14. During RAM mode operations, EnRAM is used to turn on the pass transistors so that BL0 and BL1 are electrically connected. The write driver can directly drive both bit-lines and SL during RAM mode write operations. During RAM mode read operations, current from the read bias generator flows through the pass transistors and the selected bit-cell to SL. As a result, a voltage appears on the positive input of the sense amplifier. The value of this voltage depends on the resistance of the selected bit-cell. The sense amplifier compares the voltage at its positive input to a reference voltage and outputs a '0' if the reference voltage is higher. Otherwise, the sense amplifier outputs a '1'. During ROM mode read operations, EnRAM is deasserted to turn off the pass transistors. The latch is turned on to determine which bit-line is the high impedance node. When the latch is turned on, there is a current path from BL0 to V_{DD} through M1–M4, and a current path from BL1 to V_{DD} through M1 and M6–M8 [see Fig. 7.13]. Due to the cross-coupled inverter action in the latch, the BL that is the high impedance node will get charged to V_{DD} while the other BL is discharged to GND. During ROM read operation of the scenario shown in Fig. 7.15, BL0 is discharged to GND and ROMOut outputs a '0'. If the selected bit-cell is connected to BL1 instead, BL0 is charged to V_{DD} and ROMOut outputs a '1'. Since only one of BL0 or BL1 has a direct path to GND through a SSC, a minimum sized latch may be used as the sense amplifier for ROM mode read operations. Hence, the area overhead of the peripheral circuitry may be significantly lower than that in the design in [83].

7.2.2 Evaluating ROM-embedded STT-MRAM on-chip caches

Due to the lack of suitable benchmark programs, custom benchmark programs were developed to evaluate the effectiveness of R-MRAM and R-CPSTT for evaluation of complex math functions. The benchmark programs simulate repeated calls to two commonly used math functions Sin and Log. Three steps are generally needed in

the evaluation of complex math functions using Intel's math library [82,84]: 1) range reduction, 2) approximation, and 3) reconstruction. A power series is evaluated in Step 2 to approximate the result of the function evaluation. A look-up table is used in Step 3 and combined with the result from Step 2 to obtain the accurate result of the function evaluation. Evaluation of the approximating polynomial and looking up data in the table may be executed in parallel. To achieve certain accuracy in the result of the function evaluation, the degree of the approximating polynomial used needs to be high if the size of the look-up table is small. The degree of the approximating polynomial may be reduced by increasing the size of the look-up table. The evaluation latency may be dominated by either the latency of table look-up or the latency of evaluating the approximating polynomial. If the table is stored off-chip, a small (large) table takes a shorter (longer) time to be loaded into on-chip cache. As was shown in [82], the evaluation latency can be large if the degree of the polynomial used for Step 2 is high (since it takes longer to evaluate the polynomial) or the look-up table used in Step 3 is large (since the chances of cache miss is high). To investigate the tradeoffs between the approximating polynomial and the size of the look-up table, evaluation of complex math functions was considered using look-up tables of two different sizes (2 KB and 128 KB) and use approximating polynomials of appropriate complexity to ensure similar accuracy for both scenarios. The approximation step uses a polynomial of degree 7 (degree 4) when the table size is 2 KB (128 KB). Inputs and outputs of the functions are IEEE double precision floating point numbers with at least 65 b accuracy. The average latency for each function evaluation is determined and used as the metric for the effectiveness of R-MRAM and R-CPSTT.

Layout comparisons

In order to compare R-MRAM and R-CPSTT at iso- bit-cell area, their layouts are used to determine the size of access transistor (ATx) in R-MRAM and the sizes of ATx's in R-CPSTT. Several layouts for R-MRAM and R-CPSTT, drawn using

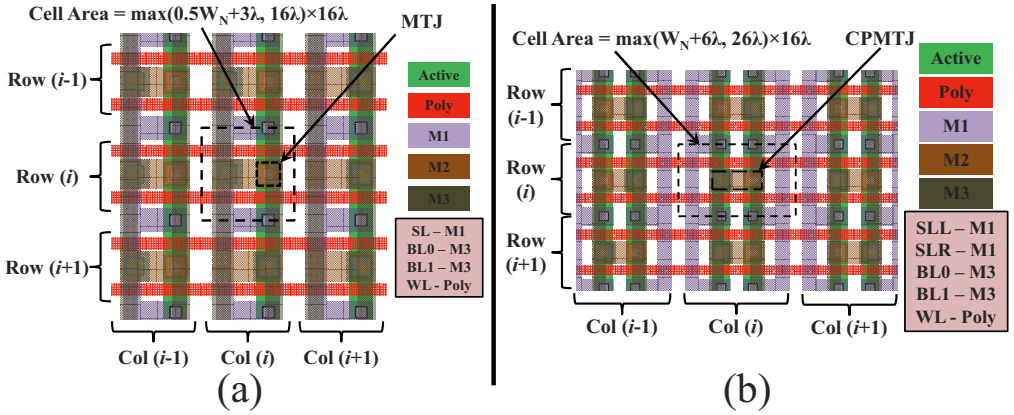


Fig. 7.16. Array layout of (a) R-MRAM and (b) R-CPSTT.

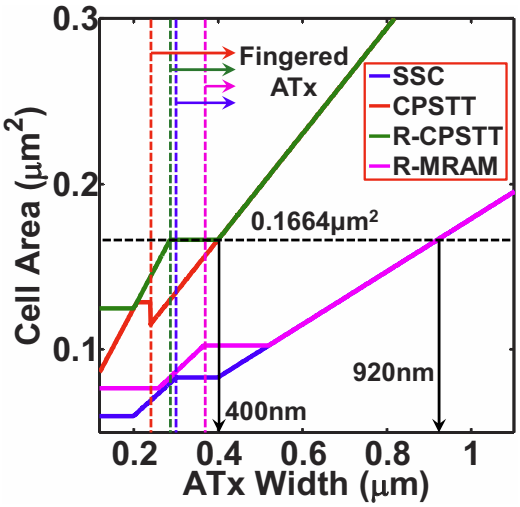


Fig. 7.17. Bit-cell area versus access transistor (ATx) width of SSC, CPSTT, R-MRAM and R-CPSTT. Vertical lines denote when the layout transitions to one using multi-finger ATx's. The bit-cell area does not change with ATx width when the layout is limited by contact or metal pitch.

λ based layout rules [69], were explored and Fig. 7.16 shows the R-MRAM and R-CPSTT layouts used for our comparisons in the rest of this paper. The bit-cell area versus ATx width for R-MRAM and R-CPSTT are plotted in Fig. 7.17. Comparisons with Standard STT-MRAM bit-Cell (SSC) and CPSTT show that ROM may be embedded without bit-cell area penalty if the ATx is large (Fig. 7.17). For the

Table 7.1.
Bit-cell Simulation Parameters

Retention Barrier Height	$56k_BT$
Write pulse width	2.0 ns
FL size (SSC)	10 nm × 10 nm × 1.5 nm
FL size (CPSTT)	10 nm × 22.5 nm × 1.5 nm
$TMR,\ RA_P$ at $V_{MTJ} = 0V$	160%, 7.5 Ω-μm^2 at $t_{MgO} = 1.15$ nm
Bit-cell Area (SSC and CPSTT)	0.1152 μm^2
ATx Width (SSC, CPSTT)	600 nm, 240 nm
t_{MgO}	1.0 nm
CMOS Technology	45 nm bulk
SSC: I_C('0'), I_C('1')	8 μA, 16 μA
CPSTT: I_C('0'), I_C('1')	13.5 μA, 13.5 μA

I_C obtained from 1300 OOMMF monodomain simulations at $T = 300$ K

following comparisons between R-MRAM and R-CPSTT, the bit-cell area is fixed at 0.1664 μm^2. The corresponding ATx widths in R-MRAM and in R-CPSTT are shown in Table 7.1.

RAM Mode Performance Evaluation

The RAM mode read performance of R-MRAM and R-CPSTT depends on the sensing scheme used. Since a self-referenced differential sensing scheme can be used for R-CPSTT but not for R-MRAM, comparison of RAM mode read performance is done using a D.C. current sensing scheme for both R-MRAM and R-CPSTT. For the RAM mode read operation of R-MRAM, a fixed read voltage (V_{READ}) is applied across the bit-cell and the read current flowing through it (I_{READ}) is compared to a reference current, I_{REF}. I_{REF} is the average of $I_{READ,L}$ (I_{READ} when the bit-cell stores a low resistance state or '0') and $I_{READ,H}$ (I_{READ} when the bit-cell stores a high resistance state or '1'). The sense amplifier outputs '1' when $I_{READ} < I_{REF}$, and '0' when $I_{READ} > I_{REF}$. For the RAM mode read operation of R-CPSTT, V_{READ} is applied to both source lines (as shown earlier in Fig. 7.5), while the bit-lines are grounded. The sense amplifier compares the I_{READ} flowing through SLL and through SLR. When

Table 7.2.
Iso-V_{READ} Comparison of Sensing Margins at $V_{DD} = 1.0V$, 2 ns Read Cycle

$V_{READ} = 0.3$ V	R-MRAM	R-CPSTT
I_{REF}	9.62 μA	6.48 μA
$I_{READ,P}$	12.46 μA	12.14 μA
$I_{READ,AP}$	6.63 μA	6.48 μA
Margin	31.1%	87.4%
Avg. Read Energy / Bit	11.55 fJ	5.59 fJ

Table 7.3.
Iso-V_{READ} Comparison of Disturb Margins at $V_{DD} = 1.0V$, 2 ns Read Cycle

R-MRAM	R-CPSTT
21.13%	50.45%

Table 7.4.
Iso-Write Margin V_{DD} and Average Write Power / Bit

Write Margin	SSC	CPSTT
0%	0.678 V, 11.35 μW	0.566 V, 11.57μW
5%	0.697 V, 12.16 μW	0.586 V, 12.51μW
10%	0.716 V, 12.99 μW	0.604 V,13.48μW
15%	0.734 V, 13.83 μW	0.623 V, 14.45μW

I_{READ} through SLL is higher (lower) than I_{READ} through SLR, the sense amplifier outputs a '0' ('1'). Note that in R-CPSTT, the bit-cell stores '0' if the resistances between BL and SLL and between BL and SLR are low and high, respectively. The bit-cell stores '1' instead if the resistances between BL and SLL and between BL and SLR are high and low, respectively. These are the only configurations possible in R-CPSTT since the free layer is parallel to only one of the two pinned layers at any time. Hence, the sensing margin for R-MRAM is defined as $\frac{min(|I_{READ,L}-I_{REF}|,|I_{READ,H}-I_{REF}|)}{I_{REF}}$, whereas it is defined as $\frac{|I_{READ,L}-I_{READ,H}|}{min(I_{READ,L},I_{READ,H})}$ for R-CPSTT. The sensing margins of R-MRAM and R-CPSTT are compared in Table 7.2. Read energy per bit of R-MRAM is 107% higher than in R-CPSTT because I_{REF} needs to be generated separately. Note that data stored in the bit-cell may be accidentally overwritten because I_{READ}

Table 7.5.
Architectural Simulation Parameters

Processor core	Out-of-order, RUU size-16 Decode width-4, Issue width-4
Functional units	Integer - 8 ALUs, 4 Multipliers Floating point - 2 ALUs, 2 Multipliers
L1 D/I cache	32KB, directly-mapped, 32B line size
L2 unified cache	2MB, 4-way associative, 64B line size

*B = Byte, K = Kilo, M = Mega

is flowing through the bit-cell during read operation, resulting in *read disturb failure*. Read disturb failures are minimized by ensuring that there is sufficient disturb margin (defined as $\frac{I_C - I_{READ}}{I_C}$). Note that the direction of I_{READ} is fixed and hence, only one type of disturb failure can occur – a stored '0' being overwritten or a stored '1' being overwritten – during read operations. Table 7.3 compares the disturb margins of R-MRAM and R-CPSTT. Furthermore, HSPICE [37] simulations performed to evaluate the read performance of R-CPSTT using a latch for sensing RAM data (like in Fig. 6.8) show that read operations up to 1.7GHz are possible.

As explained earlier in Chapter 6, the FL in R-CPSTT needs to be enlarged so as to interface with both pinned layers, resulting in a larger I_C('0') compared to R-MRAM as shown in Table 7.1. However, R-MRAM requires bi-directional write current flow to program the bit-cells in RAM mode, whereas R-CPSTT always parallelizes the free layer with a pinned layer. Hence, I_C('1') of R-CPSTT can be lower than that of R-MRAM, as shown in Table 7.1. Furthermore, the ATx's are never source degenerated during R-CPSTT RAM mode write operations. Hence, the V_{DD} for R-CPSTT to meet the required write margins (defined as write margin = $\frac{I_{WRITE} - I_C}{I_C}$, where I_{WRITE} is the current flowing through the bit-cell during write operation) can be substantially lower than that in R-MRAM to meet the same write margin. This is shown in Table 7.4. Note that the average I_{WRITE} is higher in R-CPSTT than in R-MRAM although V_{DD} is lower. Hence, the average write power per bit may be higher in R-CPSTT than in R-MRAM.

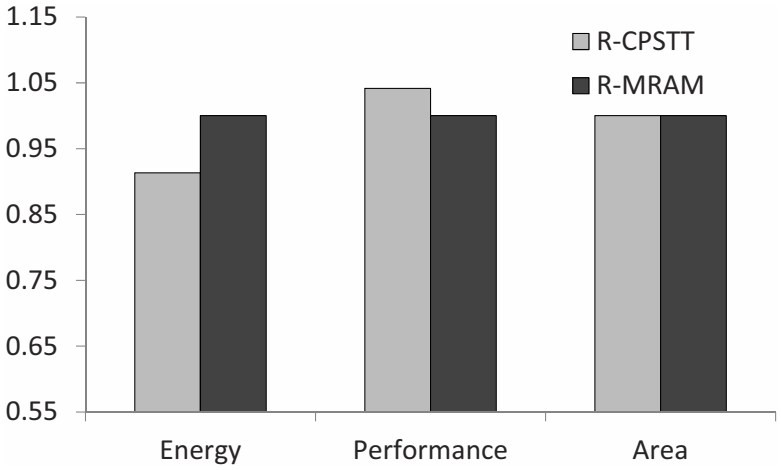


Fig. 7.18. RAM mode comparisons of R-MRAM and R-CPSTT at the architecture-level.

Since the comparison of energy consumption at the bit-cell level does not account for the fact that read operations are more frequent than write operations in many cache applications, a system level simulation was done to compare the RAM mode performance of R-MRAM and R-CPSTT. Table 7.5 shows the processor configuration used to evaluate R-MRAM and R-CPSTT in the SimpleScalar architectural simulator [79] for a wide range of SPEC2K6 benchmarks. Fig. 7.18 shows simulation results, which are normalized to R-MRAM results, for 2MB L2 cache based on R-CPSTT and on R-MRAM. R-CPSTT based L2 cache achieved 4% improvement in performance at 9% lower energy consumption as compared to R-MRAM L2 cache.

7.2.3 ROM Mode Performance Evaluation

Fig. 7.19 shows the total evaluation latency of $sin(x)$ and $log(x)$ using the conventional SRAM cache architecture (Conv), R-MRAM and R-CPSTT (normalized to the total evaluation latency using Conv) when 2KB look-up table is used. As the number of function calls increases, there is initially an increase in the improvement in performance relative to Conv case. Initial accesses to the look-up table results in cache misses in the Conv case. Therefore, a larger fraction of execution time is dominated

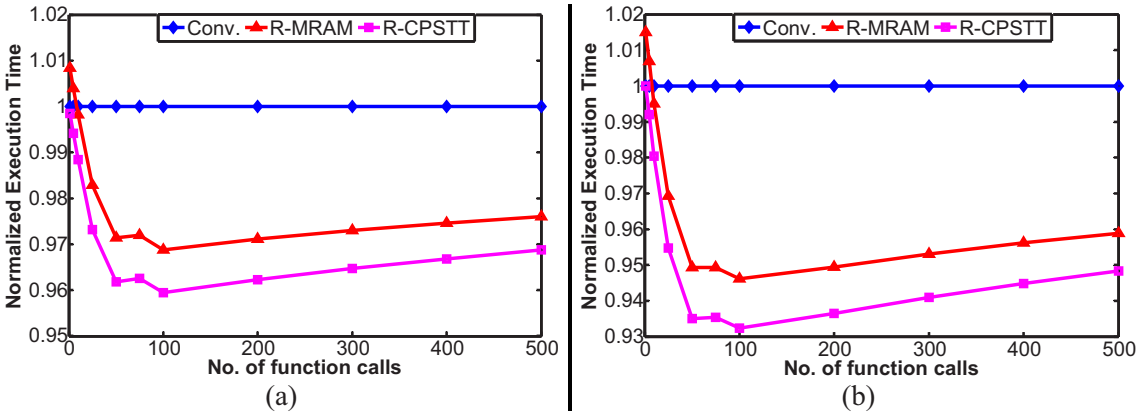


Fig. 7.19. Comparisons of evaluation latencies of (a) $log(x)$ and (b) $sin(x)$ using conventional SRAM cache (Conv.), R-MRAM, and R-CPSTT using 2KB look-up tables. R-MRAM read latency is assumed to be twice that of SRAM and R-CPSTT.

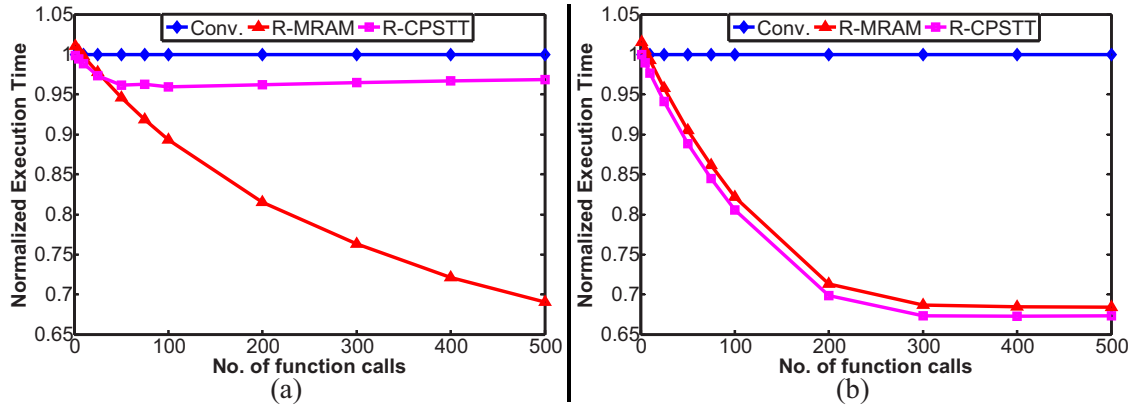


Fig. 7.20. Comparisons of evaluation latencies of (a) $log(x)$ and (b) $sin(x)$ using conventional SRAM cache (Conv.), R-MRAM, and R-CPSTT using 128KB look-up tables. R-MRAM read latency is assumed to be twice that of SRAM and R-CPSTT.

by accesses to the look-up table from memory when the number of function calls is small. As a result, increasing the number of function calls leads to large increases in execution time. However, further increase in the number of function calls increases the likelihood that the table data is completely loaded into L1 cache in the Conv case. Hence, the improvement of R-MRAM and R-CPSTT over Conv decreases when the number of function calls is more than 100. The improvements using R-MRAM over

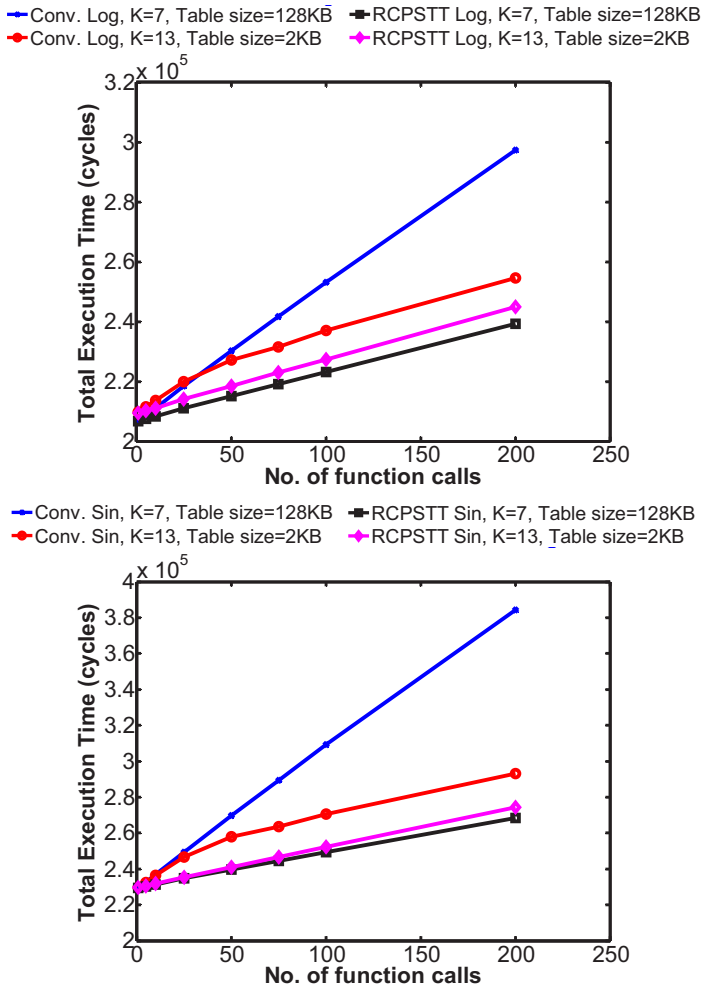


Fig. 7.21. Comparison of the total evaluation cycles for (top) $log(x)$ and (bottom) $sin(x)$ using different table sizes (and hence, approximating polynomial) to achieve 65b accuracy.

Conv are $\sim 3\%$ and $\sim 5\%$ in evaluating $log(x)$ and $sin(x)$, respectively, while the improvements using R-CPSTT over Conv in evaluating $log(x)$ and $sin(x)$ are $\sim 4\%$ and $\sim 7\%$, respectively.

Note that the evaluation latency is dominated by the latency of evaluating the approximating polynomial when 2KB look-up tables are used. Hence, the degree of the approximating polynomial may be reduced to reduce evaluation latency [82]. However, the total evaluation latency may become limited by cache read latency if the degree of the approximating polynomial is too low. Fig. 7.20 compares the

improvement in performance while using 128KB look-up tables for $sin(x)$ and $log(x)$. As shown in Fig. 7.20, R-MRAM and R-CPSTT can achieve more than $\sim$ 30% improvement in performance. The improvements remain high for a large number of function calls because the look-up table is not entirely in L1 cache. The inputs to the functions are random enough that some of the required entries of the look-up table may have been moved out of L1 cache in the Conv case and need to be reloaded.

Fig. 7.21 shows the sensitivity of Conv and R-CPSTT to the size of the look-up table with increasing number of function calls for $log(x)$ (top) and $sin(x)$ (bottom) evaluation, respectively. In the Conv case, a small look-up table leads to lower execution times because a large look-up table requires a large number of off-chip memory accesses. On the other hand, in R-CPSTT case, the performance is optimal while using a larger look-up table size. The latency of table look-up is equal to the read latency of L2 cache in R-CPSTT. Thus, the performance sensitivity to look-up table accesses in R-CPSTT is reduced. Furthermore, the degree of the approximating polynomial is small which reduces the processor workload and furthers improve performance. Note also that the execution time of R-CPSTT design is lower than Conv design for look-up table of size 2KB as well as 128KB, demonstrating the optimality of the proposed design.

7.3 Summary

The earlier chapters in this dissertation proposed the complementary polarizers MTJ (CPMTJ) structure that may significantly improve the performance of STT-MRAM for on-chip cache applications. However, the attractiveness of STT-MRAM as a candidate for future universal memory technology goes beyond the replacement of 6T SRAM in on-chip caches. This chapter showed that STT-MRAM allows the embedding of new functionality in on-chip cache almost for free and used two examples to illustrate the case: spin dice for security applications and ROM-embedded STT-MRAM on-chip cache for accelerating applications that use look-up tables. The

CPMTJ based spin dice (CPSD) consumes 14 fJ/bit during sensing and an architecture was proposed to ensure the randomness of the random number generated in the presence of process variations. The proposed ROM-embedded STT-MRAM (RSTT-MRAM) on-chip caches was used to accelerate the evaluation of complex math functions. Simulation results using the evaluation of complex math functions as an example presented in Section 7.2.3 show that the proposed RSTT-MRAM was able to reduce total execution time by as much as 30%. Although the aforementioned examples may also be embedded in 6T SRAM based caches to yield similar improvements, embedding these functions in 6T SRAM may be non-trivial and may incur more overhead as compared to in STT-MRAM based cache. Furthermore, it is possible to embed ROM and spin dice functions into the STT-MRAM cache with area overhead in terms of the control circuitry to enable the use of both functions. However, there may be conflicting design requirements for each function. For example, to ensure that the CPMTJ based STT-MRAM (CPSTT) may function as RAM, its activation energy, E_A, needs to be sufficiently high to ensure thermal stability. This is in contrast with the design requirements of the energy-efficient CPSD proposed in Section 7.1. To incorporate the spin dice function in RSTT-MRAM, the conventional architecture (with initialize, roll and sense operations) is required. Hence, a detailed analysis needs to be done to optimize RSTT-MRAM embedded with spin dice function.

8. CONCLUSION

The objective of this dissertation is to identify the design issues in spin-transfer torque magnetic RAM (STT-MRAM), to propose design techniques to overcome these design issues, and to propose design methodologies that exploit STT-MRAMs for enabling on-chip applications. The basic design of STT-MRAM was discussed and a devices-to-circuits simulation framework was proposed to evaluate STT-MRAM bit-cells. The proposed simulation framework models physical phenomena in STT-MRAM and physical parameters in the model may be used to calibrate the simulation framework to experimentally measured data.

A failure analysis methodology for STT-MRAM is then proposed. The failures in STT-MRAM are write failure, read-disturb failure and read-decision failure. The proposed methodology estimates the probability of each failure mechanism using the calibrated simulation framework proposed in this dissertation. It was shown that write failure may be severe and hence, write failure mitigation techniques were proposed. In the STT-MRAM bit-cells analyzed in this dissertation, it was found that word-line voltage boosting and applied external magnetic field may be effective in reducing write failure as well as reducing average write energy per bit.

It was then shown that the critical design issues (shared read and write current paths, source degeneration of the access transistor during write operations, and single-ended sensing of stored data) arise due to the two-terminal nature of the magnetic tunnel junction (MTJ) which is used as the storage device in conventional STT-MRAM. Hence, a multi-terminal MTJ structure consisting of complementary polarizers (the CPMTJ structure) was proposed. STT-MRAM based on the CPMTJ structure (called CPSTT), as compared to conventional STT-MRAM, can achieve 14% and 51% savings in write and read energy per bit, respectively. Furthermore, CPSTT enables self-referenced differential sensing that can achieve 6× faster read

operations than conventional STT-MRAM. The improved performance in CPSTT makes it more suitable for on-chip cache application and system-level evaluation of CPSTT based L2 cache shows that it enabled 9% improvement in processor performance as well as 9% savings in total energy consumption.

Finally, this dissertation shows that the attractiveness of STT-MRAM goes beyond on-chip cache. Design techniques for two applications–true random number generator and read-only memory embedded on-chip cache–were proposed and analyzed. These applications may be implemented in STT-MRAM based on-chip caches without any penalty (in terms of bit-cell area or cache performance). However, the complexity of the control circuitry is increased. Furthermore, several applications may be enabled simultaneously by applying the design techniques proposed (also without penalty in bit-cell area or performance) at the expense of increased complexity of the control circuitry.

9. FUTURE WORK

9.1 STT-MRAM Array Level Failure Mitigation Techniques

The failure analysis model and mitigation methodologies neglected the fact that array level failure mitigation techniques (such as adding redundant rows and columns, and implementing error correction codes or ECC) may be implemented in the design STT-MRAM. Consider for example the lowering of the activation energy, E_A, of STT-MRAM to reduce the critical write current, I_C, and hence the write power, which may be significantly higher than in high performance SRAMs [76]. However, the retention time of the STT-MRAM is reduced, which may lead to retention failures. If it can be guaranteed that the retention failure rate is sufficiently low, ECC schemes may be implemented in the array to recover from retention errors. On the other hand, the analysis of array level failure mitigation techniques already proposed in the literature do not consider the failure characteristics of the STT-MRAM memory cell at the device or circuit level [65,85]. In real STT-MRAM arrays, the additional parity bits for implementing ECC may need to be stored as well, leading to area overhead in terms of the additional bit-cells, encoder and decoder required, and performance penalty in terms of the additional delay required to encode and decode data into the code words that are stored in the STT-MRAM array. Hence, the analysis of array level failure mitigation techniques may not be accurate if accurate modeling of the STT-MRAM device and bit-cells are not included in the same analysis. Also, the array level analysis should also consider the fact that some failures may not be functionally catastrophic. Consider for example decision failures caused by stuck-at faults due to variations in the resistance of the magnetic tunnel junction (MTJ) in the STT-MRAM bit-cell. If the fault is a stuck-at-'0' and the data being written into the memory cell is a '0', the memory array is still storing the correct bit even though

the same memory cell is unable to store a '1'. Hence, the ECC scheme implemented needs to ensure that the bit being written into STT-MRAM bit-cells with stuck-at faults correspond to the stuck-at value. Another important consideration is that some of the failures in STT-MRAM are highly correlated. Take for example the memory cells with stuck-at-'1' faults, in which the MTJ has unusally high resistance due to process variations. The write current that can flow through the same bit-cell is also likely to be limited and hence, the bit-cell is more susceptible to write failures. Hence, a complete failure analysis and failure mitigation methodology that fully incorporates device-circuit-array level co-design techniques is needed to explore the possibilities enabled by implementing failure mitigation strategies at several levels of design abstraction.

9.2 Embedding New Functionality in STT-MRAM Arrays

As discussed in Chapter 6, multi-terminal storage devices may be required to overcome the design issues in STT-MRAM. However, multiple access transistors are needed to implement STT-MRAM bit-cells that use these multi-terminal storage devices, which degrades the achievable integration density. This disadvantage may be significantly offset if many additional functionality may be implemented in the same STT-MRAM array. The key idea here is that although the STT-MRAM array size may not be the smallest, the total area used to implement the memory as well as the newly embedded functions may be smaller than if each of the functions are implemented in separate circuit blocks. Two examples have been presented in Chapter 7: 1) on-chip true random number generator (TRNG) for security applications, and 2) embedded read-only memory (ROM) for accelerating specific applications.

A Physically Unclonable Function (PUF) is a security primitive that is used for secure transactions between devices [86, 87]. The memory cells in an STT-MRAM array have different measured electrical resistances due to process variations, even if all of them are storing the same data. Furthermore, the memory cell at a particular

memory address may also have different resistances depending on which die it is on. Hence, the absolute and relative resistances of the STT-MRAM bit-cells are unique to each die. As such, the comparisons of STT-MRAM bit-cell resistances with each other on the same die may be used to generate chip-unique identifiers for secure chip transactions, which corresponds to the functionality of a memory PUF, which is a *weak* PUF [86, 87]. Weak PUFs are so called because the number of possible combinations input-output pairs are small [88]. On the other hand, there are *strong* PUFs in which the number of input-output pairs can be extremely large with very complicated mappings [88]. An example of a strong PUF is an arbiter PUF, in which the signal propagation delay is also exploited to generate input-output pairs.

The multi-terminal STT-MRAM bit-cells proposed and analyzed in this dissertation may also be used to implement PUFs. It can be expected that the unique characteristics of these STT-MRAM bit-cells may be exploited to yield better PUF designs. Hence, there is a need to explore new PUFs designed using STT-MRAM bit-cells based on different multi-terminal storage devices.

LIST OF REFERENCES

LIST OF REFERENCES

[1] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits: a design perspective*. Pearson Education, 2003.

[2] S. Borkar and A. A. Chien, "The future of microprocessors," *Communications of the ACM*, vol. 54, no. 5, p. 67, May 2011.

[3] N. A. Kurd, S. Bhamidipati, C. Mozak, J. L. Miller, P. Mosalikanti, T. M. Wilson, A. M. El-Husseini, M. Neidengard, R. E. Aly, M. Nemani, M. Chowdhury, and R. Kumar, "A Family of 32 nm IA Processors," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 119–130, Jan. 2011.

[4] R. J. Riedlinger, R. Bhatia, L. Biro, B. Bowhill, E. Fetzer, P. Gronowski, and T. Grutkowski, "A 32nm 3.1 billion transistor 12-wide-issue Itanium processor for mission-critical servers," in *2011 IEEE International Solid-State Circuits Conference*. IEEE, Feb. 2011, pp. 84–86.

[5] T. Fischer, S. Arekapudi, E. Busta, C. Dietz, M. Golden, S. Hilker, A. Horiuchi, K. A. Hurd, D. Johnson, H. McIntyre, S. Naffziger, J. Vinh, J. White, and K. Wilcox, "Design solutions for the Bulldozer 32nm SOI 2-core processor module in an 8-core CPU," in *2011 IEEE International Solid-State Circuits Conference*. IEEE, Feb. 2011, pp. 78–80.

[6] S. Narenda, L. C. Fujino, and K. C. Smith, "Through the Looking Glass Continued (III): Update to Trends in Solid-State Circuits and Systems from ISSCC 2014 [ISSCC Trends]," *IEEE Solid-State Circuits Magazine*, vol. 6, no. 1, pp. 49–53, 2014.

[7] J. L. Hennessy and D. A. Patterson, *Computer Architecture, Fifth Edition: A Quantitative Approach*, 5th ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[8] 2011. [Online]. Available: http://www.isscc.org/doc/2011/2011_Trends.pdf

[9] K. C. Smith, A. Wang, and L. C. Fujino, "Through the Looking Glass?Part 2 of 2: Trend Tracking for ISSCC 2013 [ISSCC Trends]," *IEEE Solid-State Circuits Magazine*, vol. 5, no. 2, pp. 33–43, Jan. 2013.

[10] R. Keller, D. Kramer, and J.-P. Weiss, *Facing the multicore-challenge: aspects of new paradigms and technologies in parallel computing*. New York, NY, USA: Springer, 2010.

[11] 2010. [Online]. Available: http://www.itrs.net/Links/2010ITRS/2010Update/ToPost/ERD_ERM_2010FINALReportMemoryAssessment_ITRS.pdf

[12] Y. Huai, "Spin-transfer torque MRAM (STT-MRAM): challenges and prospects," *AAPPS Bulletin*, vol. 18, no. 6, pp. 33–40, 2008.

[13] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, "Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions." *Nature materials*, vol. 3, no. 12, pp. 868–71, Dec. 2004.

[14] Y. Huai, F. Albert, P. Nguyen, M. Pakala, and T. Valet, "Observation of spin-transfer switching in deep submicron-sized and low-resistance magnetic tunnel junctions," *Applied Physics Letters*, vol. 84, no. 16, p. 3118, 2004.

[15] P. Krzysteczko, X. Kou, K. Rott, A. Thomas, and G. Reiss, "Current induced resistance change of magnetic tunnel junctions with ultra-thin MgO tunnel barriers," *Journal of Magnetism and Magnetic Materials*, vol. 321, no. 3, pp. 144–147, Feb. 2009.

[16] Y. Huai, M. Pakala, Z. Diao, D. Apalkov, Y. Ding, and A. Panchula, "Spin-transfer switching in MgO magnetic tunnel junction nanostructures," *Journal of Magnetism and Magnetic Materials*, vol. 304, no. 1, pp. 88–92, Sep. 2006.

[17] D. D. Sayeef Salahuddin and S. Datta, "Spin transfer torque as a non-conservative pseudo-field," 2008.

[18] J. Slonczewski, "Current-driven excitation of magnetic multilayers," *Journal of Magnetism and Magnetic Materials*, vol. 159, no. 1-2, pp. L1–L7, Jun. 1996.

[19] L. Berger, "Emission of spin waves by a magnetic multilayer traversed by a current," *Physical Review B*, vol. 54, no. 13, pp. 9353–9358, Oct. 1996.

[20] T. Kawahara, R. Takemura, K. Miura, J. Hayakawa, S. Ikeda, Y. Lee, R. Sasaki, Y. Goto, K. Ito, T. Meguro, F. Matsukura, H. Takahashi, H. Matsuoka, and H. Ohno, "2Mb Spin-Transfer Torque RAM (SPRAM) with Bit-by-Bit Bidirectional Current Write and Parallelizing-Direction Current Read," in *2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*. IEEE, Feb. 2007, pp. 480–617.

[21] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu, M. Nowak, and N. Yu, "45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell," in *2009 IEEE International Electron Devices Meeting (IEDM)*. IEEE, Dec. 2009, pp. 1–4.

[22] R. Nebashi, N. Sakimura, H. Honjo, S. Saito, Y. Ito, S. Miura, Y. Kato, K. Mori, Y. Ozaki, Y. Kobayashi, N. Ohshima, K. Kinoshita, T. Suzuki, K. Nagahara, N. Ishiwata, K. Suemitsu, S. Fukami, H. Hada, T. Sugibayashi, and N. Kasai, "A 90nm 12ns 32Mb 2T1MTJ MRAM," in *2009 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*. IEEE, Feb. 2009, pp. 462–463,463a.

[23] 2009. [Online]. Available: http://www.itrs.net/Links/2009ITRS/2009Chapters_2009Tables/2009_PIDS.pdf

[24] J. Sun, "Current-driven magnetic switching in manganite trilayer junctions," *Journal of Magnetism and Magnetic Materials*, vol. 202, no. 1, pp. 157–162, Jun. 1999.

[25] X. Wang, W. Zhu, and D. Dimitrov, "Quantum transport and stochastic magnetization dynamics simulation on intrinsic spin torque switching asymmetry," *Physical Review B*, vol. 79, no. 10, pp. 1–5, Mar. 2009.

[26] D. Apalkov, Z. Diao, A. Panchula, S. Wang, Y. Huai, and K. Kawabata, "Temperature Dependence of Spin Transfer Switching in Nanosecond Regime," *IEEE Transactions on Magnetics*, vol. 42, no. 10, pp. 2685–2687, Oct. 2006.

[27] Y. Huai, M. Pakala, Z. Diao, Y. Ding, D. Apalkov, and A. Panchula, "Spin transfer switching and spin polarization in magnetic tunnel junctions with MgO and AlO_x barriers," *Applied Physics Letters*, vol. 87, no. 23, p. 232502, 2005.

[28] X. Yao, H. Meng, Y. Zhang, and J.-P. Wang, "Improved current switching symmetry of magnetic tunneling junction and giant magnetoresistance devices with nano-current-channel structure," *Journal of Applied Physics*, vol. 103, no. 7, p. 07A717, 2008.

[29] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, and H. Ohno, "A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction." *Nature materials*, vol. 9, no. 9, pp. 721–4, Sep. 2010.

[30] G. Jeong, W. Cho, S. Ahn, H. Jeong, G. Koh, Y. Hwang, and K. Kim, "A 0.24μm 2.0V 1T1MTJ 16kb NV magnetoresistance RAM with self reference sensing," in *2003 IEEE International Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC.*, vol. 1. IEEE, 2003, pp. 280–281.

[31] Y. Chen, H. Li, X. Wang, and W. Zhu, "A nondestructive self-reference scheme for Spin-Transfer Torque Random Access Memory (STT-RAM)," in *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, no. c. IEEE, Mar. 2010, pp. 148–153.

[32] J.-B. Kammerer, M. Madec, and L. Hébrard, "Compact Modeling of a Magnetic Tunnel JunctionPart I: Dynamic Magnetization Model," *IEEE Transactions on Electron Devices*, vol. 57, no. 6, pp. 1408–1415, Jun. 2010.

[33] M. Madec, J.-B. Kammerer, and L. Hébrard, "Compact Modeling of a Magnetic Tunnel JunctionPart II: Tunneling Current Model," *IEEE Transactions on Electron Devices*, vol. 57, no. 6, pp. 1416–1424, Jun. 2010.

[34] S. Lee, H. Lee, S. Kim, S. Lee, and H. Shin, "A novel macro-model for spin-transfer-torque based magnetic-tunnel-junction elements," *Solid-State Electronics*, vol. 54, no. 4, pp. 497–503, Apr. 2010.

[35] J. D. Harms, F. Ebrahimi, X. Yao, and J.-p. Wang, "SPICE Macromodel of Spin-Torque-Transfer-Operated Magnetic Tunnel Junctions," *IEEE Transactions on Electron Devices*, vol. 57, no. 6, pp. 1425–1430, Jun. 2010.

[36] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, S. Takahashi, M. Nakayama, N. Shimomura, H. Aikawa, S. Ikegawa, S. Yuasa, K. Yakushiji, H. Kubota, A. Fukushima, M. Oogane, T. Miyazaki, and K. Ando, "Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM," in *2008 IEEE International Electron Devices Meeting*. IEEE, Dec. 2008, pp. 1–4.

[37] "HSPICE." [Online]. Available: http://www.synopsys.com/Tools/Verification/AMSVerification/CircuitSimulation/HSPICE/

[38] J. Li, P. Ndai, A. Goel, S. Salahuddin, and K. Roy, "Design Paradigm for Robust Spin-Torque Transfer Magnetic RAM (STT MRAM) From Circuit/Architecture Perspective," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 12, pp. 1710–1723, Dec. 2010.

[39] W. Xu, H. Sun, X. Wang, Y. Chen, and T. Zhang, "Design of Last-Level On-Chip Cache Using Spin-Torque Transfer RAM (STT RAM)," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 3, pp. 483–493, Mar. 2011.

[40] C.-K. Koh, W.-F. Wong, Y. Chen, and H. Li, "The salvage cache: A fault-tolerant cache architecture for next-generation memory technologies," in *2009 IEEE International Conference on Computer Design*. IEEE, Oct. 2009, pp. 268–274.

[41] H. H. Li and Y. Chen, "Emerging non-volatile memory technologies: From materials, to device, circuit, and architecture," in *2010 53rd IEEE International Midwest Symposium on Circuits and Systems*. IEEE, Aug. 2010, pp. 1–4.

[42] J. Li, S. Salahuddin, and K. Roy, "Variation-tolerant Spin-Torque Transfer (STT) MRAM array for yield enhancement," in *2008 IEEE Custom Integrated Circuits Conference*. IEEE, Sep. 2008, pp. 193–196.

[43] J. Li, C. Augustine, S. Salahuddin, and K. Roy, "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement," in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*. New York, New York, USA: IEEE, Jun. 2008, pp. 278–283.

[44] N. N. Mojumder, S. K. Gupta, S. H. Choday, D. E. Nikonov, and K. Roy, "A Three-Terminal Dual-Pillar STT-MRAM for High-Performance Robust Memory Applications," *IEEE Transactions on Electron Devices*, vol. 58, no. 5, pp. 1508–1516, May 2011.

[45] N. N. Mojumder and K. Roy, "Proposal for Switching Current Reduction Using Reference Layer With Tilted Magnetic Anisotropy in Magnetic Tunnel Junctions for Spin-Transfer Torque (STT) MRAM," *IEEE Transactions on Electron Devices*, vol. 59, no. 11, pp. 3054–3060, Nov. 2012.

[46] N. N. Mojumder, S. K. Gupta, and K. Roy, "Dual Pillar Spin Transfer Torque MRAM with tilted magnetic anisotropy for fast and error-free switching and near-disturb-free read operations," in *69th Device Research Conference*, vol. 54, no. 2010. IEEE, Jun. 2011, pp. 67–68.

[47] S. Fukami, T. Suzuki, K. Nagahara, N. Ohshima, Y. Ozaki, S. Saito, R. Nebashi, N. Sakimura, H. Honjo, K. Mori, C. Igarashi, S. Miura, N. Ishiwata, and T. Sugibayashi, "Low-Current Perpendicular Domain Wall Motion Cell for Scalable High-Speed MRAM," in *2009 Symposium on VLSI Technology*, Jun. 2009, pp. 230–231.

[48] S. Datta, *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, 1997.

[49] M. D'Aquino, "Nonlinear magnetization dynamics in thin-films and nanoparticles," Ph.D. dissertation, University of Naples Federico II, 2004.

[50] S. Salahuddin, D. Datta, P. Srivastava, and S. Datta, "Quantum Transport Simulation of Tunneling Based Spin Torque Transfer (STT) Devices: Design Trade offs and Torque Efficiency," in *2007 IEEE International Electron Devices Meeting*. IEEE, 2007, pp. 121–124.

[51] D. Datta, B. Behin-Aein, S. Datta, and S. Salahuddin, "Voltage Asymmetry of Spin-Transfer Torques," *IEEE Transactions on Nanotechnology*, vol. 11, no. 2, pp. 261–272, Mar. 2012.

[52] T. Shima, K. Takanashi, Y. K. Takahashi, and K. Hono, "Coercivity exceeding 100 kOe in epitaxially grown FePt sputtered films," *Applied Physics Letters*, vol. 85, no. 13, p. 2571, 2004.

[53] N. N. Mojumder, C. Augustine, D. E. Nikonov, and K. Roy, "Effect of quantum confinement on spin transport and magnetization dynamics in dual barrier spin transfer torque magnetic tunnel junctions," *Journal of Applied Physics*, vol. 108, no. 10, p. 104306, 2010.

[54] X. Fong, S. H. Choday, G. Panagopoulos, C. Augustine, and K. Roy, "SPICE Models for Magnetic Tunnel Junctions Based on Monodomain Approximation," Aug. 2013. [Online]. Available: https://nanohub.org/resources/19048

[55] "Object Oriented MicroMagnetic Framework." [Online]. Available: http://math.nist.gov/oommf/software-12a4pre.html

[56] G. Fuchs, J. Katine, S. Kiselev, D. Mauri, K. Wooley, D. Ralph, and R. Buhrman, "Spin Torque, Tunnel-Current Spin Polarization, and Magnetoresistance in MgO Magnetic Tunnel Junctions," *Physical Review Letters*, vol. 96, no. 18, pp. 1–4, May 2006.

[57] W. C. Jeong, J. H. Park, J. H. Oh, G. H. Koh, G. T. Jeong, H. S. Jeong, and K. Kim, "Field assisted spin switching in magnetic random access memory," *Journal of Applied Physics*, vol. 99, no. 8, p. 08H708, 2006.

[58] T. Devolder, P. Crozat, J.-V. Kim, C. Chappert, K. Ito, J. a. Katine, and M. J. Carey, "Magnetization switching by spin torque using subnanosecond current pulses assisted by hard axis magnetic fields," *Applied Physics Letters*, vol. 88, no. 15, p. 152502, 2006.

[59] W. Zhao, T. Devolder, Y. Lakys, J. Klein, C. Chappert, and P. Mazoyer, "Design considerations and strategies for high-reliable STT-MRAM," *Microelectronics Reliability*, vol. 51, no. 9-11, pp. 1454–1458, Sep. 2011.

[60] R. Takemura, T. Kawahara, K. Ono, K. Miura, H. Matsuoka, and H. Ohno, "Highly-scalable disruptive reading scheme for Gb-scale SPRAM and beyond," *2010 IEEE International Memory Workshop*, pp. 1–2, 2010.

[61] K. Ito, T. Devolder, C. Chappert, M. J. Carey, and J. A. Katine, "Micromagnetic simulation on effect of oersted field and hard axis field in spin transfer torque switching," *Journal of Physics D: Applied Physics*, vol. 40, no. 5, pp. 1261–1267, Mar. 2007.

[62] ——, "Micromagnetic simulation of spin transfer torque switching combined with precessional motion from a hard axis magnetic field," *Applied Physics Letters*, vol. 89, no. 25, p. 252509, 2006.

[63] K. Shimura, N. Ohshima, S. Miura, R. Nebashi, T. Suzuki, H. Hada, S. Tahara, H. Aikawa, T. Ueda, T. Kajiyama, and H. Yoda, "Magnetic and Writing Properties of Clad Lines in a Toggle MRAM," in *INTERMAG 2006 - IEEE International Magnetics Conference*. IEEE, May 2006, pp. 733–733.

[64] W. J. Gallagher and S. S. P. Parkin, "Development of the magnetic tunnel junction MRAM at IBM: From first junctions to a 16-Mb MRAM demonstrator chip," *IBM Journal of Research and Development*, vol. 50, no. 1, pp. 5–23, Jan. 2006.

[65] W. Xu, Y. Chen, X. Wang, and T. Zhang, "Improving STT MRAM Storage Density through Smaller-Than-Worst-Case Transistor Sizing," in *Design Automation Conference 2009*, 2009, pp. 87–90.

[66] M. Y. Zhuravlev, Y. Wang, S. Maekawa, and E. Y. Tsymbal, "Tunneling electroresistance in ferroelectric tunnel junctions with a composite barrier," *Applied Physics Letters*, vol. 95, no. 5, p. 052902, 2009.

[67] S. Sivasubramanian, A. Widom, and Y. Srivastava, "Equivalent circuit and simulations for the Landau-Khalatnikov model of ferroelectric hysteresis." *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 50, no. 8, pp. 950–7, Aug. 2003.

[68] J. Xiao, A. Zangwill, and M. Stiles, "Boltzmann test of Slonczewskis theory of spin-transfer torque," *Physical Review B*, vol. 70, no. 17, p. 172405, Nov. 2004.

[69] C. Mead and L. Conway, *Introduction to VLSI Systems*. Addison-Wesley, 1980.

[70] "MOSIS Scalable CMOS (SCMOS) Design Rules." [Online]. Available: http://www.mosis.com/pages/design/rules/index

[71] A. D. Kent, B. Ozyilmaz, and E. del Barco, "Spin-transfer-induced precessional magnetization reversal," *Appl. Phys. Lett.*, vol. 84, no. 19, p. 3897, Apr. 2004.

[72] T. Seki, S. Mitani, K. Yakushiji, and K. Takanashi, "Magnetization reversal by spin-transfer torque in 90 configuration with a perpendicular spin polarizer," *Appl. Phys. Lett.*, vol. 89, no. 17, p. 172504, Oct. 2006.

[73] R. Sbiaa, S. Y. H. Lua, R. Law, H. Meng, R. Lye, and H. K. Tan, "Reduction of switching current by spin transfer torque effect in perpendicular anisotropy magnetoresistive devices (invited)," *J. Appl. Phys.*, vol. 109, no. 7, p. 07C707, 2011.

[74] M. Iijima, M. Kitamura, M. Numa, A. Tada, and T. Ipposhi, "Ultra Low Voltage Operation with Bootstrap Scheme for Single Power Supply SOI-SRAM," in *20th International Conference on VLSI Design held jointly with 6th International Conference on Embedded Systems (VLSID'07)*. IEEE, 2007, pp. 609–614.

[75] D. A. Patterson and J. L. Hennessy, *Computer Organization and Design, Revised Fourth Edition: The Hardware/Software Interface*. Elsevier, 2011, vol. 2011.

[76] S. P. Park, S. Gupta, N. Mojumder, A. Raghunathan, and K. Roy, "Future cache design using STT MRAMs for improved energy efficiency," in *Proceedings of the 49th Annual Design Automation Conference on - DAC '12*. New York, New York, USA: ACM Press, 2012, p. 492.

[77] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA Organizations and Wiring Alternatives for Large Caches with CACTI 6.0," in *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*. IEEE, Dec. 2007, pp. 3–14.

[78] S. P. Park, S. Y. Kim, D. Lee, J.-J. Kim, W. P. Griffin, and K. Roy, "Column-selection-enabled 8T SRAM array with 1R/1W multi-port operation for DVFS-enabled processors," in *IEEE/ACM International Symposium on Low Power Electronics and Design*, ser. ISLPED '11. Piscataway, NJ, USA: IEEE, Aug. 2011, pp. 303–308.

[79] T. Austin, E. Larson, and D. Ernst, "SimpleScalar: an infrastructure for computer system modeling," *Computer*, vol. 35, no. 2, pp. 59–67, 2002.

[80] A. Fukushima, H. Kubota, K. Yakushiji, S. Yuasa, and K. Ando, "United States Patent: 8521795," 2013.

[81] N. D. Rizzo, M. DeHerrera, J. Janesky, B. Engel, J. Slaughter, and S. Tehrani, "Thermally activated magnetization reversal in submicron magnetic tunnel junctions for magnetoresistive random access memory," *Applied Physics Letters*, vol. 80, no. 13, p. 2335, 2002.

[82] D. Lee and K. Roy, "Area Efficient ROM-Embedded SRAM Cache," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 9, pp. 1583–1595, Sep. 2013.

[83] D. Lee, X. Fong, and K. Roy, "R-MRAM: A ROM-Embedded STT MRAM Cache," *IEEE Electron Device Letters*, vol. 34, no. 10, pp. 1256–1258, Oct. 2013.

[84] J. Harrison, T. Kubaska, S. Story, and P. T. Tang, "The Computation of Transcendental Functions on the IA-64 Architecture," *Intel Technology Journal*, vol. Q4, pp. 1–7, 1999.

[85] H. Naeimi, C. Augustine, A. Raychowdhury, S.-l. Lu, and J. Tschanz, "STTRAM Scaling and Retention Failure," *Intel Technology Journal*, vol. 17, no. 1, pp. 54–75, 2013.

[86] L. Zhang, X. Fong, C.-H. Chang, Z. H. Kong, and K. Roy, "Feasibility study of emerging non-volatilememory based physical unclonable functions," in *2014 IEEE 6th International Memory Workshop (IMW)*. IEEE, May 2014, pp. 1–4.

[87] L. Zhang, X. Fong, C.-h. Chang, Z. H. Kong, and K. Roy, "Highly reliable memory-based Physical Unclonable Function using Spin-Transfer Torque MRAM," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, Jun. 2014, pp. 2169–2172.

[88] U. Rührmair, F. Sehnke, J. Sölter, G. Dror, S. Devadas, and J. Schmidhuber, "Modeling attacks on physical unclonable functions," in *Proceedings of the 17th ACM conference on Computer and communications security - CCS '10*. New York, New York, USA: ACM Press, 2010, p. 237.

[89] A. Newell, W. Williams, and D. Dunlop, "A generalization of the demagnetizing tensor for nonuniform magnetization," *Journal of geophysical research*, vol. 98, no. B6, pp. 9551–9555, 1993.

[90] W. Brown, "Thermal Fluctuations of a Single-Domain Particle," *Physical Review*, vol. 130, no. 5, pp. 1677–1686, Jun. 1963.

[91] W. Scholz, "Micromagnetic simulation of thermally activated switching in fine particles," Ph.D. dissertation, Vienna University of Technology, 1999.

[92] S. Zhang and Z. Li, "Roles of Nonequilibrium Conduction Electrons on the Magnetization Dynamics of Ferromagnets," *Physical Review Letters*, vol. 93, no. 12, p. 127204, Sep. 2004.

[93] P. Braganca, J. Katine, N. Emley, D. Mauri, J. Childress, P. Rice, E. Delenia, D. Ralph, and R. Buhrman, "A Three-Terminal Approach to Developing Spin-Torque Written Magnetic Random Access Memory Cells," *IEEE Transactions on Nanotechnology*, vol. 8, no. 2, pp. 190–195, Mar. 2009.

APPENDICES

A. NON-EQUILIBRIUM GREEN'S FUNCTION BASED MTJ MODEL

The Non-Equilibrium Green's Function (NEGF) based transport model was proposed in [50, 51] and is repeated here for completeness. The NEFG model is based on the single band effective mass Hamiltonian ($\mathcal{H}$) and self-energy ($\Sigma_{L,R}$) which are used to calculate Green's function (G), electron correlation matrix (G^n) and charge current density (J). Fig. A.1 shows the device structure and coordinate system used for modeling MTJs. For each transverse mode, the Hamiltonian is written as

$$\textbf{Left Contact: } \mathcal{H}_L(i,j) = \begin{cases} \alpha_{HL1} + \left(\frac{qV}{2} + \left(\frac{I - \vec{\sigma}\cdot\widehat{M}}{2}\right)\Delta\right) I, \, i = j \\ -t_{FM} I, \, i \text{ and } j \text{ are nearest neighbors} \\ 0, \text{ otherwise} \end{cases} \tag{A.1}$$

$$\textbf{Oxide Channel: } \mathcal{H}_{OX}(i,j) = \begin{cases} \alpha_{OX} + \left(U_b + qV\left(\frac{1}{2} - \frac{i}{N+1}\right)\right) I, \, i = j \\ -t_{OX} I, \, i \text{ and } j \text{ are nearest neighbors} \\ 0, \text{ otherwise} \end{cases} \tag{A.2}$$

$$\textbf{Right Contact: } \mathcal{H}_R(i,j) = \begin{cases} \alpha_{HL2} + \left(\frac{qV}{2} + \left(\frac{I - \vec{\sigma}\cdot\widehat{m}}{2}\right)\Delta\right) I, \, i = j \\ -t_{FM} I, \, i \text{ and } j \text{ are nearest neighbors} \\ 0, \text{ otherwise} \end{cases} \tag{A.3}$$

where each 2×2 entry in $\mathcal{H}$ describes the coupling between the i-th and j-th lattice site in Fig. A.1 (i.e., $\mathcal{H}$ is a $(2N + 8) \times (2N + 8)$ matrix and N is the number of lattice sites in the oxide channel). I is the 2×2 identity matrix, $\vec{\sigma}$ represents the Pauli matrices, $\widehat{m}$ is the unit vector representing the magnetization of the right contact, $\widehat{M}$ is the unit vector representing the magnetization of the left contact, U_b is the barrier height of oxide relative to the equilibrium Fermi level in the contacts

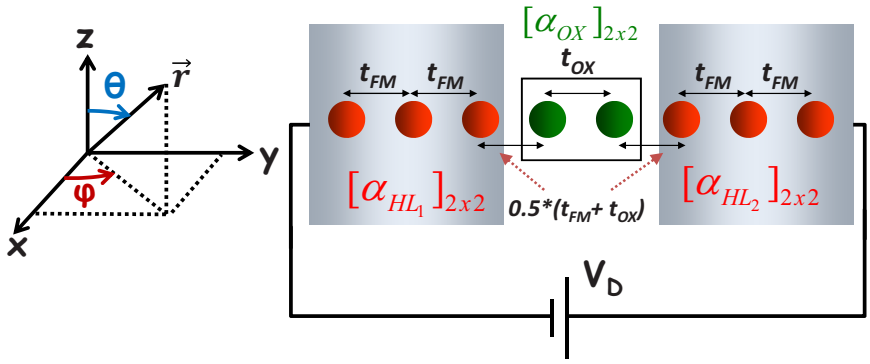


Fig. A.1. Illustration of the reference axis (left) and Non-Equilibrium Green's Function based description of the magnetic tunnel junction. The coupling between lattice sites are t_{FM} and t_{OX} and individual lattice sites are described by the Hamiltonian α_{HL1}, α_{HL2} and α_{OX}. The complete Hamiltonian describing the MTJ is written in terms of t_{FM}, t_{OX}, α_{HL1}, α_{HL2} and α_{OX}.

(E_F), and Δ is the spin splitting. For each wave vector, k_t, corresponding to each transverse mode, $\alpha_{OX}(k_t) = \left(2t_{OX} + \frac{\hbar^2 k_t^2}{2m^*_{OX}}\right) I$ and $\alpha_{FM}(k_t) = \left(2t_{FM} + \frac{\hbar^2 k_t^2}{2m^*_{FM}}\right) I$, where $t_{FM} = \frac{\hbar^2}{2m^*_{FM}a^2}$; $t_{OX} = \frac{\hbar^2}{2m^*_{OX}a^2}$; and a is the uniform lattice spacing. The coupling at the interfaces is given by

$$t_{interface} = 0.5\,(t_{OX} + t_{FM}) \tag{A.4}$$

The self-energy matrices, $\Sigma_{L,R}$, represent the coupling of the external system to the contacts and its non-zero components may be written as

$$\Sigma_{L,R}(i,i) = \begin{bmatrix} -t_{FM} exp\left(-ik^{\uparrow}_{L,R}a\right) & 0 \\ 0 & -t_{FM} exp\left(-ik^{\downarrow}_{L,R}a\right) \end{bmatrix} \tag{A.5}$$

where

$$k^{\uparrow}_{L,R} = cos^{-1}\left(1 - \frac{E \pm \frac{qV}{2} - \frac{\hbar^2 k_t^2}{2m^*_{FM}}}{2t_{FM}}\right) \tag{A.6}$$

$$k^{\downarrow}_{L,R} = cos^{-1}\left(1 - \frac{E \pm \frac{qV}{2} - \frac{\hbar^2 k_t^2}{2m^*_{FM}} - \Delta}{2t_{FM}}\right) \tag{A.7}$$

Use $+\frac{qV}{2}$ for the left contact, and $-\frac{qV}{2}$ for the right contact. E is the energy level of interest. This form of Eq. A.5 is used if the quantization axis for the spin is in the z axis. A unitary transformation is done if the quantization axis for the spin is not in the z-axis. The matrix for unitary transformation is given by

$$B_{trans} = \begin{bmatrix} cos\left(\frac{\theta}{2}\right) exp\left(i\frac{\phi}{2}\right) & sin\left(\frac{\theta}{2}\right) exp\left(-i\frac{\phi}{2}\right) \\ -sin\left(\frac{\theta}{2}\right) exp\left(i\frac{\phi}{2}\right) & cos\left(\frac{\theta}{2}\right) exp\left(-i\frac{\phi}{2}\right) \end{bmatrix} \tag{A.8}$$

where θ and ϕ correspond to the relative angles of the magnetizations to the reference axis, and the self-energy matrices are modified using

$$\Sigma_{L,R}^{new} = B_{trans}\Sigma_{L,R}^{old}B_{trans}^{\dagger} \tag{A.9}$$

With the Hamiltonian ($\mathcal{H}$) and the self-energy matrices ($\Sigma_{L,R}$), all quantities of interest may be calculated from the following:

$$\textbf{Green's function: } G(E) = (EI - \mathcal{H} - \Sigma_L - \Sigma_R)^{-1} \tag{A.10}$$

$$\textbf{Spectral density: } A = i(G - G^{\dagger}) = G\Gamma G^{\dagger} \tag{A.11}$$

$$\textbf{Electron correlation function: } G^n(E) = G(\Sigma_L^{in} + \Sigma_R^{in})G^{\dagger} \tag{A.12}$$

$$\textbf{In-scattering function: } \Sigma_{L,R}^{in}(E) = \Gamma_{L,R}(E)f_{L,R}(E) \tag{A.13}$$

$$\textbf{Broadening matrix: } \Gamma_{L,R}(E) = i(\Sigma_{L,R} - \Sigma_{L,R}^{\dagger}) \tag{A.14}$$

The diagonal elements of A and G^n correspond to the local density of states and electron density, respectively. Σ^{in} is the in-scattering function describing the rate at which electrons enter the device from the L and R contacts, and $f_{L,R}(E)$ are the Fermi functions in the L and R contacts.

A.1 Solution of MTJ currents using mode space calculations in NEGF

The charge and spin currents in the MTJ can be calculated using

Charge current density

$$J_{k,k+1} = Real\left(\frac{1}{i\hbar}\int_E \left(Trace\left(\mathcal{H}_{k,k+1}G_{k+1,k}^n - G_{k,k+1}^n\mathcal{H}_{k+1,k}\right)\right)dE\right) \tag{A.15}$$

Spin current density

$$\begin{aligned}\vec{J}_S &= J^{Spin}_{k,k+1} \\ &= Real\left(\frac{1}{i\hbar}\int_E \left(Trace\left\{\widehat{\sigma}\cdot\left(\mathcal{H}_{k,k+1}G^n_{k+1,k} - G^n_{k,k+1}\mathcal{H}_{k+1,k}\right)\right\}\right)dE\right)\end{aligned} \tag{A.16}$$

where G^n is the electron correlation function. The charge and spin currents calculated using the NEGF approach are used to determine the spin-transfer torque exerted on the free ferromagnetic layer of the MTJ. The calculation of the spin-transfer torque is discussed in detail in Appendix C.

B. MICROMAGNETICS AND MAGNETIZATION DYNAMICS IN MTJ

As mentioned in Chapter 1, data is stored as the magnetic configuration of the MTJ. Hence, the transient simulation of any 1T-1MTJ STT-MRAM bit-cell requires the transient simulation of the free layer magnetization. Depending on the size of the free layer, it may be modeled as a single magnetic particle (a single ferromagnetic domain, also called the macro-spin approximation) or as an ensemble of magnetic particles (multi-domain). The magnetization dynamics of a single magnetic particle is described by the Landau-Lifshitz-Gilbert (LLG) equation [49] which is written as

$$\frac{\partial \widehat{m}}{\partial t} = -|\gamma| \widehat{m} \times \vec{H}_{EFF} + \alpha \widehat{m} \times \frac{\partial \widehat{m}}{\partial t} + \overrightarrow{Torque} \tag{B.1}$$

where $\vec{H}_{EFF}$ is the effective magnetic field that the magnetic particle sees; $\widehat{m} = \frac{\vec{M}}{M_S}$ is the unit vector pointing in the direction of the magnetization of the magnetic particle (M_S and $\vec{M}$ are the saturation magnetization and magnetization vector of the particle, respectively); and $\overrightarrow{Torque}$ represents the sum of other torques acting on the particle. Any torque that does not come from magnetic field-like phenomenon may enter as $\overrightarrow{Torque}$ in the LLG equation. The time evolution of the magnetization of the particle can be obtained by numerically integrating the LLG equation.

Since it is often easier to work with the explicit form of the LLG instead of the implicit form shown in Eq. B.1, the LLG equation may be rewritten as

$$\frac{1+\alpha^2}{\gamma} \frac{\partial \widehat{m}}{\partial t} = -\widehat{m} \times \vec{H}_{EFF} - \alpha \widehat{m} \times \widehat{m} \times \vec{H}_{EFF} + \frac{1}{|\gamma|} \left(\alpha \widehat{m} \times \overrightarrow{Torque} + \overrightarrow{Torque} \right) \tag{B.2}$$

Eqs. B.1 and B.2 are mathematically equivalent since $|\widehat{m}| = 1$. Finally, a natural time unit, $\tau = \frac{|\gamma|}{1+\alpha^2} t$, may be defined to rewrite Eq. B.2 as

$$\frac{\partial \widehat{m}}{\partial \tau} = -\widehat{m} \times \vec{H}_{EFF} - \alpha \widehat{m} \times \widehat{m} \times \vec{H}_{EFF} + \frac{1}{|\gamma|} \left(\alpha \widehat{m} \times \overrightarrow{Torque} + \overrightarrow{Torque} \right) \tag{B.3}$$

Eq. B.3 is preferred for estimating the impact of changes in $\vec{H}_{EFF}$ and $\overrightarrow{Torque}$ on magnetization dynamics of the magnetic particle.

Details of LLG have been discussed in [49] and the focus will now shift toward $\vec{H}_{EFF}$ and $\overrightarrow{Torque} = \overrightarrow{STT}$. $\overrightarrow{STT}$ is the spin-transfer torque acting on the magnetic particle due to electron flow, which will be discussed in Appendix C. The approach for determining magnetic field-like torques is to first write the free energy, U_{free}, describing the source of the torque. U_{free} depends on the magnetization of the magnetic particle and the equivalent magnetic field acting on the particle due to U_{free} is $\vec{H} = -\vec{\nabla}U_{free}$. This is repeated for all magnetic field-like torque sources that needs

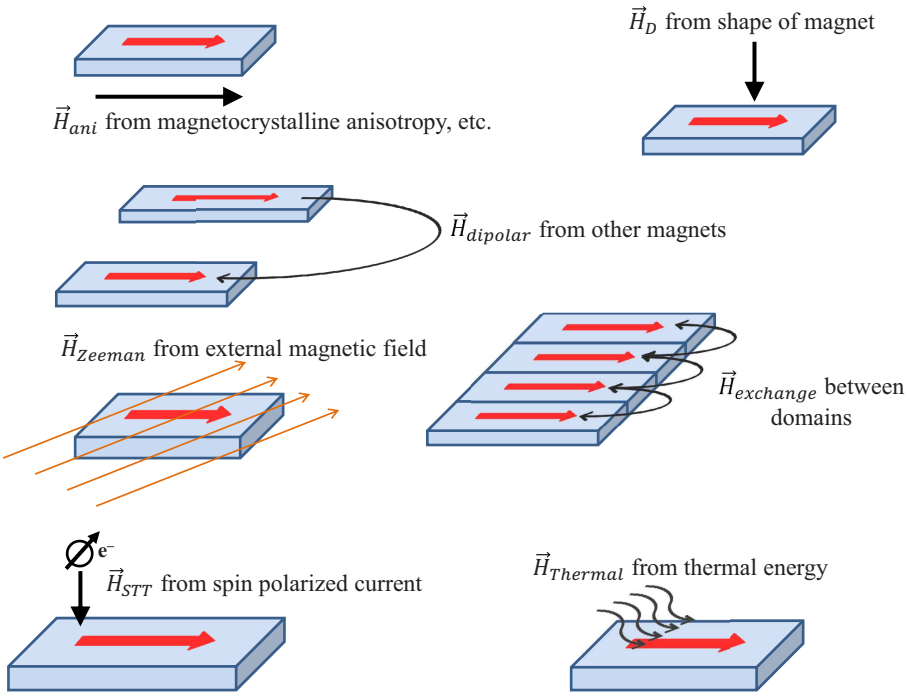


Fig. B.1. The magnetic interactions considered in this dissertation are uniaxial and cubic anisotropies (due to magnetocrystalline anisotropy, etc.), the magnetostatic or demagnetizing field giving rise to shape anisotropy, dipolar coupling with other magnets, externally applied magnetic fields, exchange interactions between magnetic domains, spin-transfer torque, and thermal fluctuations.

to be captured and the superposition of all equivalent magnetic fields is $\vec{H}_{EFF}$. In other words, write

$$U_{free} = \sum U_{magnetic-field-like-source} \tag{B.4}$$

and then

$$\vec{H}_{EFF} = -\vec{\nabla} U_{free} \tag{B.5}$$

The different magnetic interactions that this dissertation focuses on and their sources are briefly summarized in Fig. B.1. The following sections discuss the modeling of magnetic-field-like energies whereas spin-transfer torque is discussed in Appendix C.

B.1 Free Energies in a Magnet

The free energies in magnetic particles have been described in [49] and are discussed here for completeness. The free energies of interest are anisotropy energy, exchange energy, Zeeman energy, magnetostatic energy, and thermal energy.

B.1.1 Anisotropy energy

Anisotropic effects are commonly observed in ferromagnetic particles since they result from the lattice structure and certain symmetries in certain crystals. Energetically favorable directions often exist in a given magnetic material in the absence of external magnetic fields. These directions are called the *easy directions* in the literature. The free energy functional for anisotropy energy, $U_{ani}(\widehat{m})$, will have minima along the easy directions. Saddle points and maxima of $U_{ani}(\widehat{m})$ correspond to the *medium-hard* and *hard* directions, respectively.

Uniaxial anisotropy

The uniaxial anisotropy is a commonly observed anisotropy effect and corresponds to magnetic particles with only one easy direction. Hence, $U_{ani}(\widehat{m})$ will be rotationally-symmetric with respect to the easy axis. Consequently, $U_{ani}(\widehat{m})$ must

depend on the relative orientation of $\widehat{m}$ with respect to the easy axis. Suppose that a ferromagnetic particle has its easy axis along the z-axis, θ is the angle between $\widehat{m}$ and the z-axis, and ϕ as the counterclockwise angle between the $+x$-direction and the projection of $\widehat{m}$ onto the x-y plane. Due to the single axis of rotational symmetry, the Taylor series expansion of $U_{ani}(\widehat{m})$ may be written as

$$U_{ani}(\widehat{m}) = K_0 + \sum_{i=1}^{\infty} K_i (sin\ \theta)^{2i} \tag{B.6}$$

where all K_0 and K_i's are anisotropy constants having dimensions of energy per unit volume [J/m^3]. Terms that are higher than second order may be ignored. Hence, the anisotropy energy functional may be rewritten as

$$U_{ani}(\widehat{m}) = K_0 + K_1 (sin\ \theta)^2 \tag{B.7}$$

The anisotropic behavior of the ferromagnetic particle depends on the sign of the anisotropy constant K_1. For the particle with its easy direction in the z-axis, the minima of $U_{ani}(\widehat{m})$ occur at $\theta = 0$ and at $\theta = \pi$ when $K_1 > 0$. This case is also called the *easy axis anisotropy*. Fig. B.2(a) shows how $U_{ani}(\widehat{m})$ with $K_1 > 0$ may be visualized. The value of $U_{ani}(\widehat{m})$ at any point on the surface is the distance between that point and the origin ($x = y = z = 0$). Fig. B.2(b) shows that when $K_1 < 0$, the minima of $U_{ani}(\widehat{m})$ is at $\theta = \frac{\pi}{2}$ instead. The easy direction of a magnetic particle having this anisotropy energy is in any direction in the x-y plane. Hence, this case is also called *easy plane anisotropy*.

There may be a need to simulate systems that consist of several magnets with different directions of uniaxial anisotropy. Hence, the form of $U_{ani}(\widehat{m})$ in Eq. B.7 needs to be generalized. Denoting the unit vector pointing along the anisotropy direction as $\widehat{u}$, the general form of $U_{ani}(\widehat{m})$ is given by

$$U_{ani}(\widehat{m}) = K_0 + K_1 \left(1 - (\widehat{m} \cdot \widehat{u})^2\right) \tag{B.8}$$

By setting $K_1 > 0$, the easy axis of the magnetic particle will be in the direction along $\widehat{u}$. By setting $K1 < 0$, the easy plane of the magnetic particle will be in the plane perpendicular to $\widehat{u}$.

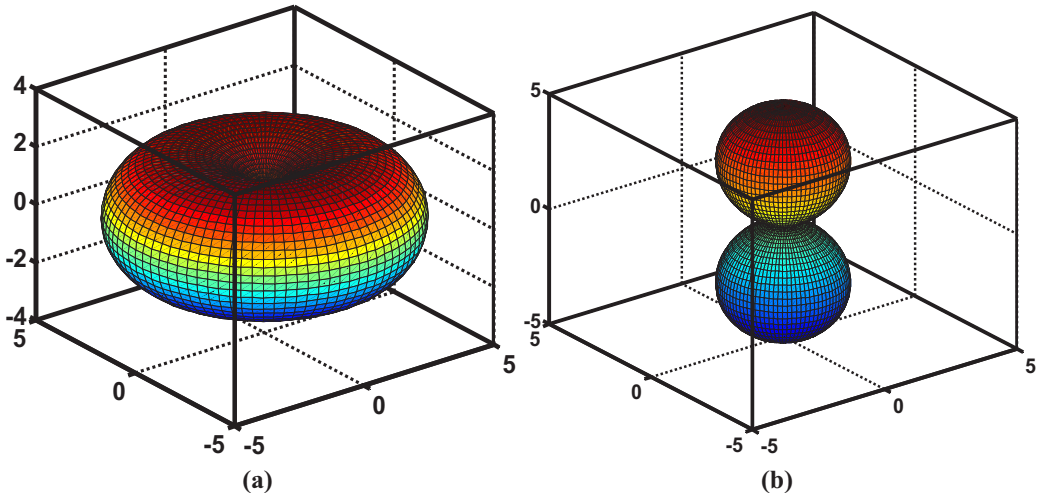


Fig. B.2. Visualizations of $U_{ani}(\widehat{m})$ for uniaxial anisotropy. (a) $K_0 = 1$ and $K_1 = 4$ results in easy axis anisotropy as indicated by the minima along z-axis. (b) $K_0 = 5$ and $K_1 = -4.5$ results in easy plane anisotropy as indicated by the minima when $m_z = 0$.

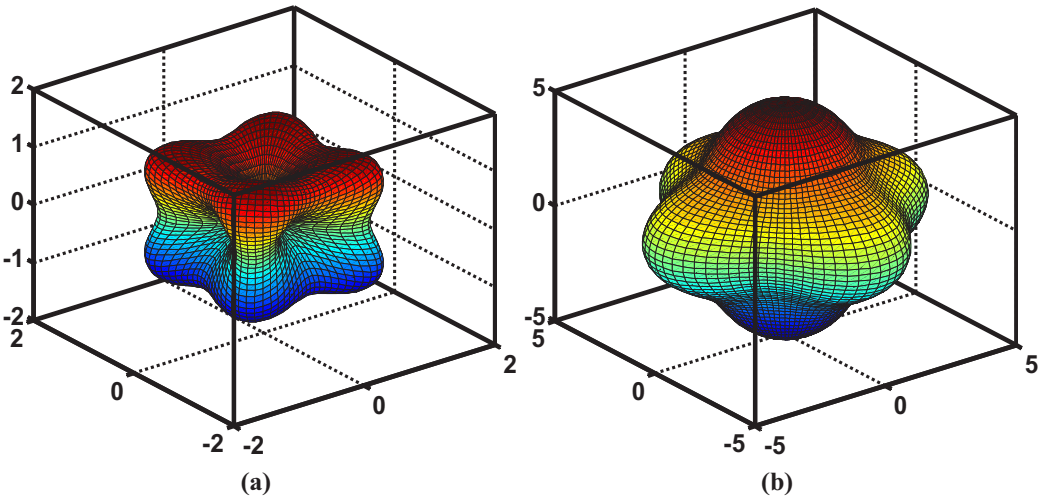


Fig. B.3. Visualizations of $U_{ani}(\widehat{m})$ for cubic anisotropy with $K_2 = 0$. (a) two minima along each of x, y and z axes (six minima in total) occur when $K_0 = 0.1$ and $K_1 = 4$. (b) When $K_0 = 5$ and $K_1 = -4.8$, two maxima along each of x, y and z axes (six maxima in total) occur.

Cubic anisotropy

Cubic anisotropy is used to describe magnetic particles in which their easy or hard directions have cubic symmetry. The anisotropy energy can be written as

$$U_{ani}(\widehat{m}) = K_0 + K_1 \left(m_x^2 m_y^2 + m_y^2 m_z^2 + m_x^2 m_z^2\right) + K_2 m_x^2 m_y^2 m_z^2 \qquad \text{(B.9)}$$

where all terms above fourth order are ignored except for $K_2 m_x^2 m_y^2 m_z^2$. For simplicity, the following discussion assumes $K_2 = 0$. As in the case for uniaxial anisotropy, the difference in $U_{ani}(\widehat{m})$ between the $K_1 > 0$ case and the $K_1 < 0$ case are investigated. Fig. B.3(a) shows the cubic anisotropy energy landscape for the case when $K_1 > 0$. There are three pairs of minima with each pair along each axial direction (x, y and z axis). However, as shown in Fig. B.3(b), three pairs of maxima are present when $K_1 < 0$, with each pair along each axial direction.

B.1.2 Exchange energy

In a large magnet composed of many smaller ferromagnetic particles, two types of exchange based phenomenon–ferromagnetism and anti-ferromagnetism–have been observed. When the ensemble behaves as a ferromagnet, all the particles in it tend to have parallel magnetization. The ensemble will then have some effective magnetization parallel to the average magnetization of all the particles. On the other hand, when the ensemble behaves as an anti-ferromagnet, neighboring particles have antiparallel magnetizations. This results in the overall ensemble having zero magnetization. The exchange energy, U_{ex}, models these effects and may be written as

$$U_{ex} = A_{ex}\left((\nabla\widehat{m})^2\right) = A_{ex}\left((\nabla m_x)^2 + (\nabla m_y)^2 + (\nabla m_z)^2\right) \tag{B.10}$$

where $\widehat{m} = \widehat{m}\,(\vec{r})$ is the spatial variation of magnetization in the ensemble. A ferromagnetic particle with uniform magnetization does not have exchange energy since $\nabla\widehat{m} = 0$.

B.1.3 Zeeman energy

The Zeeman energy corresponds to the energy due an external magnetic field acting on the magnetic particle and the energy functional, U_{Zeeman}, may be written as

$$U_{Zeeman} = -\widehat{m} \cdot \vec{H} \tag{B.11}$$

where $\vec{H}$ is the magnetic field that is acting on the magnetic particle.

B.1.4 Magnetostatic energy

Whereas the exchange energy describes nearest neighbor coupling between magnetic particles, the magnetostatic energy describes the long range coupling between magnetic particles [89]. Each magnetic particle has its own magnetic field that may extend throughout the entire 3-D space. Hence, the magnetic field of each magnetic particle in an ensemble of magnetic particles may affect the rest of the particles in the ensemble. In order to model this effect, the magnetic field due to a ferromagnetic particle everywhere in space needs to be calculated. The magnetic field due to the particle may be calculated by first noting that in a ferromagnetic body [89]

$$\vec{\nabla} \times \vec{H} = 0 \tag{B.12}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{B.13}$$

Hence, the magnetic field due to the particle is the gradient of a potential

$$\vec{H} = -\nabla \Phi_M \tag{B.14}$$

where the potential, Φ_M is calculated as

$$\begin{aligned} \Phi_M(\vec{r}) &= \frac{1}{4\pi} \int \vec{M}(\vec{r}\,') \cdot \vec{\nabla}' \left(\frac{1}{|\vec{r} - \vec{r}\,'|} \right) d\vec{r}\,' \\ &= \frac{1}{4\pi} \vec{M}' \cdot \int_{\tau'} \vec{\nabla}' \left(\frac{1}{|\vec{r} - \vec{r}\,'|} \right) d\vec{r}\,' \end{aligned} \tag{B.15}$$

where the last integration is over the region occupied by the particle and $\vec{\nabla}'$ is the gradient with respect to $\vec{r}\,'$. In a region, τ, which may overlap τ', the average magnetic field in τ is

$$\left\langle \vec{H}' \right\rangle_\tau = \frac{1}{\tau} \int_\tau \left(-\vec{\nabla} \Phi_M \right) d\tau = -\vec{M}' \cdot N(\vec{r}) \tag{B.16}$$

where $N(\vec{r})$ is a 3×3 tensor (called the *demagnetizing tensor*) at every $\vec{r}$. The components of $N(\vec{r})$ is given by

$$N_{ij} = -\frac{1}{4\pi\tau} \int_\tau d\tau \int_{\tau'} \vec{\nabla}'_i \vec{\nabla}'_j \left(\frac{1}{|\vec{r} - \vec{r}\,'|} \right) d\tau' \tag{B.17}$$

which may be transformed by Gauss's theorem into surface integrals

$$\begin{aligned} N &= \frac{1}{4\pi\tau}\int_{\tau} d\tau \vec{\nabla}\int_{\tau'}\vec{\nabla}'\left(\frac{1}{|\vec{r}-\vec{r}'|}\right)d\tau' \\ &= \frac{1}{4\pi\tau}\int_{\vec{S}} d\vec{S}\int_{\vec{S}'}\frac{d\vec{S}'}{|\vec{r}-\vec{r}'|} \end{aligned} \tag{B.18}$$

with $d\vec{S} = \widehat{n}dS$, where $\widehat{n}$ is the normal to the surface. If the ensemble of ferromagnetic particles are cuboids aligned with the Cartesian axis, the interpretation of Eqs. B.17 and B.18 is as follows. Each component of the demagnetizing tensor describes j-th component of the demagnetizing field at $\vec{r}$

$$\vec{H}_D = -N\left(\vec{r}-\vec{r}'\right)\cdot\vec{M}\left(\vec{r}'\right) \tag{B.19}$$

due to monopoles distributed on the surfaces of the source particle at $\vec{r}'$ along the i-th direction. Hence, Eq. B.16 reduces to the magnetic field due to a magnetic dipole when the ferromagnetic particles being considered are sufficiently far apart:

$$\vec{H}_{dipolar} = \frac{3\vec{R}\left(\vec{M}\cdot\vec{R}\right)-\vec{M}\left|\vec{R}\right|^2}{4\pi\left|\vec{R}\right|^5} \tag{B.20}$$

where $\vec{M} = M_S\widehat{m}$ is total magnetic moment of the source particle i which has its magnetization pointing along the unit vector $\widehat{m}$, and $\vec{R}$ is the vector pointing from particle i to the destination particle j. The reader may refer to [89] for further details regarding the computation of N.

Note that the computation of the demagnetizing tensor for a single particle is a time consuming process. In a multi-domain problem, the computation time of $\vec{H}_D$ grows as $O(n^2)$, where n is the number of particles in the problem, if the interactions between particles are considered pairwise. However, the computation of the demagnetizing field by pairwise consideration of particles may result in many redundant calculations. Note that the demagnetizing tensor depends only on $\vec{R}$, which describes the position of the destination particle with respect to the source particle. In a multi-domain problem, there may be many pairs of particles with the same relative position to each other. A careful inspection of Eqs. B.15–B.19 reveals that $\vec{H}_D$ is the result

of a convolution operation. Hence, the computation of $\vec{H}_D$ may be accelerated using fast Fourier transform (FFT). Using a source particle at the origin (i.e., $\vec{r}' = \vec{0}$), $N(\vec{r})$ is first computed over a range of $\vec{r}$. The region covered by $\vec{r}$ must be sufficient to enclose all particles in the micromagnetic problem of interest (i.e., The largest $\vec{r}$ corresponds to at least the separation between the two farthest separated particles in the ensemble). The FFT of $N(\vec{r})$ is then calculated and stored. During simulation of the micromagnetic problem, the FFT of $\vec{M}(\vec{r})$ is computed and multiplied with the FFT of $N(\vec{r})$. The inverse FFT of the result of the multiplication gives $\vec{H}_D$. The FFT method to calculate $\vec{H}_D$ is used in almost all fast micromagnetic solvers such as the Object-Oriented MicroMagnetic Framework (OOMMF) [55].

B.1.5 Thermal energy

Brown Jr. proposed a model for the effect of thermal energy on a single ferromagnetic particle [90]. The details of this model are beyond the scope of this dissertation. It is sufficient to know that the thermal energy is modeled as a Wiener process, and that the effect of thermal fluctuations can be captured using a thermally fluctuating magnetic field acting on the ferromagnetic particle. Analysis of the Fokker-Planck equation shows the thermal field has the statistical properties

$$\left\langle \vec{H}_{fluct,i} \right\rangle = 0,\ i = x,\ y,\ z \tag{B.21}$$

$$\left\langle \vec{H}_{fluct,i}(t)\vec{H}_{fluct,i}(t+\tau) \right\rangle = \frac{2\alpha k_B T}{|\gamma| M_S V}\delta(\tau)\delta_{ij} \tag{B.22}$$

where α, γ and M_S are the material dependent parameters of the magnetic particle; α is the Gilbert damping factor, γ is the gyromagnetic ratio, and M_S is the saturation magnetization. V is the volume of the magnetic particle, k_B is the Boltzmann constant, and T is the absolute temperature of the magnetic particle. Hence, the LLG equation is transformed into the stochastic LLG (sLLG) equation when considering effects due to temperature, and the thermal field, $\vec{H}_{Thermal}$ is given as

$$\vec{H}_{Thermal} = \vec{\xi}\sqrt{\frac{2\alpha k_B T}{|\gamma| M_S V}} \tag{B.23}$$

During transient simulation of the magnetization dynamics with a time discretization of dt, the thermal field at any particular time is generated as

$$\vec{H}_{Thermal} = \vec{\xi}\sqrt{\frac{2\alpha k_B T}{|\gamma| M_S V dt}} \tag{B.24}$$

where the factor of dt appears to ensure the total magnetic energy is averaged to zero in the numerical solution obtained. $\vec{\xi}$ is a 3-D vector whose components are zero mean Gaussian random variables with standard deviation of 1. The reader may refer to [91] for details regarding the numerical solution to the sLLG equation, which are beyond the scope of this dissertation. Simulations used to obtain the results presented in this dissertation take into consideration the mathematical details presented in [91] to ensure artifacts of numerical simulation do not appear in the results.

C. SPIN-TRANSFER TORQUE

Spin-transfer torque was theoretically predicted by Slonczewski [18] and Berger [19] and describes the transfer of spin angular momentum from itinerant electrons incident on a ferromagnetic body. The following sections presents the spin-transfer torque modeling framework used in this dissertation. As mentioned earlier in Appendix A.1, the spin and charge current densities through an MTJ is used to calculate the spin-transfer torque acting on the free layer. Two approaches have been proposed for calculating the spin-transfer torque vector, $\overrightarrow{STT}$. The first method was proposed in [18, 19, 68], and the second was proposed in [50, 51].

C.1 Slonczewski's Formulation of Spin-Transfer Torque

The method proposed by in [18, 19, 68] has since been known as the Slonczewski spin-transfer torque theory and the proposed modification to the LLG equation (resulting in the Landau-Lifshitz-Gilbert-Slonczewski or LLGS equation) is as follows

$$\frac{\partial \widehat{m}}{\partial t} = -|\gamma|\widehat{m} \times \vec{H}_{EFF} + \alpha \widehat{m} \times \frac{\partial \widehat{m}}{\partial t} + \overrightarrow{STT} \tag{C.1}$$

$$\overrightarrow{STT} = |\gamma|\beta \left(\widehat{m} \times \left(\epsilon \widehat{m} \times \widehat{M} + \epsilon' \widehat{M} \right) \right) \tag{C.2}$$

$$\beta = \frac{\hbar J_{MTJ}}{2e\mu_0 M_S t_{FL}} \tag{C.3}$$

$$\epsilon = \left[\frac{q_+}{A_+ + A_- \left(\widehat{m} \cdot \widehat{M} \right)} + \frac{q_-}{A_+ - A_- \left(\widehat{m} \cdot \widehat{M} \right)} \right] \tag{C.4}$$

$$q_{\pm} = \left[P_{PL}\Lambda_{PL}^2 \sqrt{\frac{\Lambda_{FL}^2 + 1}{\Lambda_{PL}^2 + 1}} \pm P_{FL}\Lambda_{FL}^2 \sqrt{\frac{\Lambda_{PL}^2 - 1}{\Lambda_{FL}^2 - 1}} \right] \tag{C.5}$$

$$A_{\pm} = \sqrt{\left(\Lambda_{PL}^2 \pm 1\right)\left(\Lambda_{FL}^2 \pm 1\right)} \tag{C.6}$$

$$\Lambda^2 = GR \tag{C.7}$$

$$G = \frac{A_{MTJ} q^2 k_F^2}{4\pi^2 \hbar} \tag{C.8}$$

where P_{PL}, P_{FL}, Λ_{PL} and Λ_{FL} are fitting parameters of the model. J_{MTJ} is the charge current density flowing through the free layer, e is the electronic charge, μ_0 is the permeability of vacuum, M_S is the saturation magnetization of the ferromagnet, $\overrightarrow{M}$ is the unit vector pointing in the direction of pinned layer magnetization; and t_{FL} is the length of the current path through the free layer. For a standard MTJ with cross-sectional area, A_{MTJ}, where the current flows perpendicular to the ferromagnet-oxide-ferromagnet interfaces, $J_{MTJ} = \frac{I_{MTJ}}{A_{MTJ}}$ where I_{MTJ} is the total current flowing through the MTJ and t_{FL} is the thickness of the free layer.

Eqs. C.1–C.8 may be intuitively interpreted by considering the standard MTJ structure and rewriting Eq. C.3 as

$$\beta = \frac{\hbar}{2} \frac{I_{MTJ}}{e} \frac{1}{\mu_0 M_S V_{FL}} \epsilon \tag{C.9}$$

where $V_{FL} = A_{MTJ} \times t_{FL}$. Note that the first term on the right-hand side of Eq. C.9 is the spin-angular momentum carried by an electron. The second term describes the rate of electrons passing through the MTJ. The product of the first two terms gives the total spin angular momentum carrier by electrons flowing through the MTJ, which is also the total amount of spin angular momentum that may be transferred to the free layer. An interpretation of ϵ is that it is a dimensionless factor that describes the effectiveness of the spin-transfer process between the electrons and the free layer of the MTJ if all the electrons have identical spin directions. From the discussion in Chapter 1, it is clear that although the ferromagnetic free and pinned layers in the MTJ act as spin polarizers, they do not completely spin polarize all electrons if they are not perfect half-metals. An alternative interpretation of ϵ is that it describes

the degree to which the electrons flowing through the MTJ are spin-polarized. The fact that the relative density of states of the free and pinned layers determines the degree of spin-polarization (which was briefly described in Section 1.1) supports this alternative view–ϵ depends on $\left(\vec{m}\cdot\vec{M}\right)$, which describes the relative density of states of the free and pinned layers. This interpretation of ϵ is also supported in the NEGF formalism [50, 51] as will be explained in the later sections.

Before discussing the NEGF approach to calculating $\overrightarrow{STT}$, it is worthy to note that an alternative equation for $\overrightarrow{STT}$ found in the literature is given by [92]

$$\overrightarrow{STT} = -|\gamma|\widehat{m}\times\left(b_J\left(\widehat{m}\times\vec{\nabla}\widehat{M}\right)+c_J\vec{\nabla}\widehat{M}\right) \tag{C.10}$$

which consists of an adiabatic and a nonadiabatic term. However, the key difference between Eq. C.10 and Eqs. C.1–C.8 is the pinned layer magnetization–replacing $\overrightarrow{M}$ in Eq. C.2 with $\vec{\nabla}\widehat{M}$ yields the same form as Eq. C.10. A detailed investigation of this difference is beyond the scope of this dissertation. However, it should be noted that in the context of micromagnetic simulations, electron transport between magnetic domains needs to be considered when modeling spin-transfer torque. Consider when electrons flow from left to right in the $+x$-direction through three magnetic domains. The electrons entering the middle domain should be spin-polarized in the magnetization direction of the leftmost domain. Hence, the leftmost domain acts like a pinned layer for the middle domain. However, the electrons are leaving the middle domain out to the rightmost domain. Hence, the rightmost domain also acts like a pinned layer for the middle domain. The total spin-transfer torque, $\overrightarrow{STT'}$, on the middle domain may be calculated as

$$\overrightarrow{STT'} = \overrightarrow{STT}_{i-1} + \overrightarrow{STT}_{i+1} \tag{C.11}$$

$$\overrightarrow{STT}_{i-1} = |\gamma|\frac{\beta'}{dx}\left(\widehat{m}\times\left(\epsilon\widehat{m}\times\widehat{M}_{i-1}+\epsilon'\widehat{M}_{i-1}\right)\right) \tag{C.12}$$

$$\overrightarrow{STT}_{i+1} = -|\gamma|\frac{\beta'}{dx}\left(\widehat{m}\times\left(\epsilon\widehat{m}\times\widehat{M}_{i+1}+\epsilon'\widehat{M}_{i+1}\right)\right) \tag{C.13}$$

where $\beta' = \frac{\hbar J_{MTJ}}{2q\mu_0 M_S}$, and dx is the length of the particle along the x-direction. If these particles are cuboids in a finite difference grid for multi-domain simulation, dx is the separation between neighboring grid points in the x-direction. Then,

$$\overrightarrow{STT'} = -|\gamma|\beta' \left(\widehat{m} \times \left(\epsilon \widehat{m} \times \frac{\widehat{M_P}}{dx} + \epsilon' \frac{\widehat{M_P}}{dx} \right) \right) \tag{C.14}$$

where $\widehat{M_P} = \widehat{M_{i+1}} - \widehat{M_{i-1}}$. If dx is reduced to infinitesimally small, then

$$\lim_{dx \to 0} \frac{\widehat{M_P}}{dx} = \lim_{dx \to 0} \frac{\widehat{M_{i+1}} - \widehat{M_{i-1}}}{dx} \tag{C.15}$$

$$= \lim_{dx \to 0} \frac{\widehat{M_{i+1}} - \widehat{M_i} + \widehat{M_i} - \widehat{M_{i-1}}}{dx} \tag{C.16}$$

$$= \vec{\nabla} \widehat{M} \tag{C.17}$$

and hence

$$\overrightarrow{STT'} = -|\gamma|\beta' \left(\widehat{m} \times \left(\epsilon \widehat{m} \times \left(\vec{\nabla} \widehat{M} \right) + \epsilon' \left(\vec{\nabla} \widehat{M} \right) \right) \right) \tag{C.18}$$

Thus, Eq. C.10 and Eq. C.2 are mathematically equivalent if $\beta'\epsilon = b_J$ and $\beta'\epsilon' = c_J$.

C.2 NEGF Approach to Spin-Transfer Torque

The NEGF based approach proposed in [50, 51] uses the spin currents calculated using the NEGF formalism to directly write

$$\frac{\partial \widehat{m}}{\partial t} = -|\gamma| \widehat{m} \times \vec{H}_{EFF} + \alpha \widehat{m} \times \frac{\partial \widehat{m}}{\partial t} + \overrightarrow{STT} \tag{C.19}$$

$$\begin{aligned} \overrightarrow{STT} &= -\mu_B \int_\Omega \left(\vec{\nabla} \cdot \vec{J}_S \right) dV = \mu_B \int_S \int_y - \left(\vec{\nabla} \cdot \vec{J}_S \right) dy \, dS \\ &= \mu_B \int_S \left(\vec{J}_{S,L} - \vec{J}_{S,R} \right) dS \end{aligned} \tag{C.20}$$

where $\vec{J}_S$ is given in Eq. A.16. $\vec{J}_{S,L}$ and $\vec{J}_{S,R}$ correspond to $\vec{J}_S$ calculated at the lattice site in the free layer that is directly adjacent to the oxide and farthest from the oxide, respectively. If the spin-transfer torque is completely absorbed by the free layer, the exiting current is completely spin polarized and

$$\overrightarrow{STT} = \mu_B \int_S \left(\vec{J}_{S,L} - J_{MTJ} \widehat{m} \right) dS \tag{C.21}$$

where $\vec{\sigma}$ are the Pauli spin matrices.

D. MULTI-TERMINAL MAGNETIC TUNNEL JUNCTIONS AS STT-MRAM STORAGE DEVICES

It has been shown in this dissertation that two-terminal MTJs for STT-MRAM requires the read and write current paths to be shared, which leads to severe design limitations. Although the two-terminal nature of the storage device allows for very small bit-cell footprint, the footprint needs to be enlarged if better STT-MRAM performance is required. Several multi-terminal MTJ (MTMTJ) structures have been proposed in the literature to mitigate the design issues in STT-MRAM [44, 47, 93]. These MTMTJs alleviate the conflicting design requirements in STT-MRAM by decoupling the read and write current paths. Although MTMTJs may require an additional access transistor (ATx) in the bit-cell, the sizing requirements of the ATx may be less stringent than that in STT-MRAM based on two-terminal MTJs. Hence, STT-MRAM bit-cells using MTMTJs may have smaller footprint than those based on two-terminal MTJs. The MTMTJs proposed in the literature are reviewed in this section for completeness.

D.1 The Dual-Pillar MTJ Structure

The dual pillar MTJ (DPMTJ) structure mitigates the design issues in STT-MRAM by decoupling the read and write current paths [44, 93]. The read and the write current paths may then be optimized independently. As shown in Fig. D.1 and D.2, the DPMTJ structure consists of a free layer (FL), a pinned layer (PL) that is called the *read port*, and a PL that is called the *write port*. Data is stored as the magnetization direction of the FL, which may be sensed as the resistance of the DPMTJ through the read port. Write current (I_{WRITE}) flows through between FL and the PL on the write port during write operations, whereas read current (I_{READ})

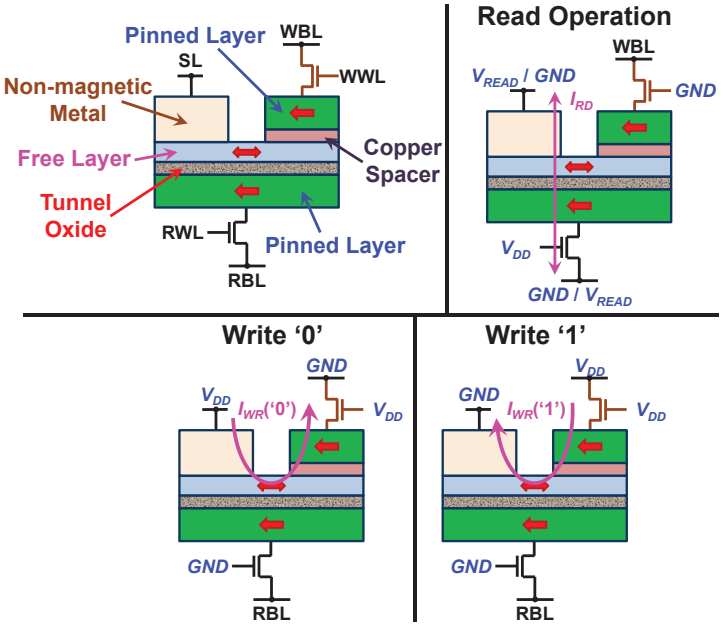


Fig. D.1. The dual pillar MTJ (DPMTJ) proposed in [93]

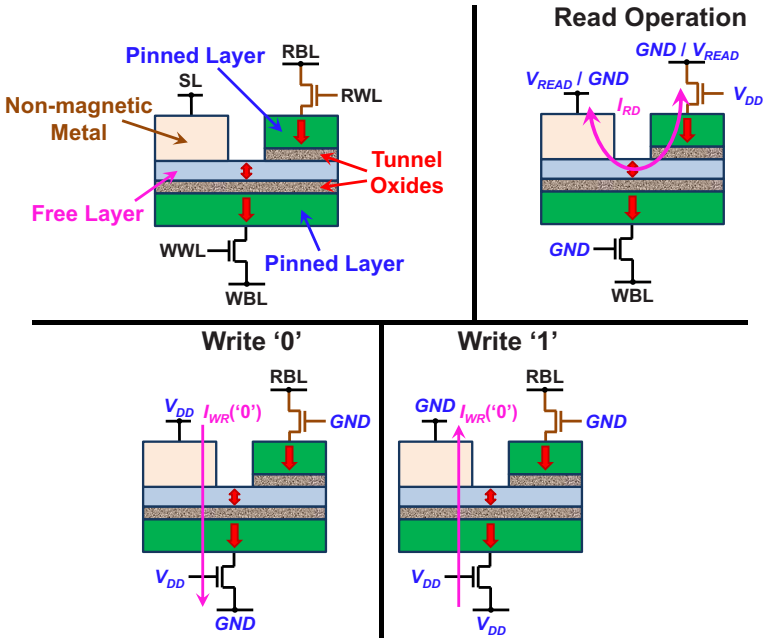


Fig. D.2. An alternative DPMTJ structure proposed in [44].

flows between FL and the PL on the read port during read operations. Note that since large write currents do not flow through the tunnel junction in the read port, the reliability of the tunnel oxide in the read port, which is crucial for the TMR and hence the readability of the tunnel junction, is improved.

The DPMTJ proposed in [93] consists of a spin-valve and a tunneling junction with a shared FL as shown in Fig. D.1. The PL is formed first and a tunneling oxide is formed on top of it. The FL is then deposited on top to form a tunnel junction. Two metallic contacts (one of which is Cu while the other is Cr/Au) are then deposited on the top of the FL. Another PL is formed on top of the Cu contact to create a spin-valve. The PL in the spin-valve of the DPMTJ is designated as the write port and the other PL is designated as the read port. Write operations occur by passing I_{WRITE} through the low resistance path between the Cr/Au electrode and the PL in the spin-valve. Read operations occur by passing I_{READ} through the high resistance path between the Cr/Au electrode and the tunnel junction instead.

Since the DPMTJ in [93] uses a spin-valve in the write port, the spin polarization efficiency of I_{WRITE} may be degraded. Furthermore, the large cross-sectional area of the read port reduces the absolute difference between resistance states and may degrade the distinguishability of the stored states. Hence, an alternative DPMTJ structure was proposed in [44] as shown in Fig. D.2. The difference between the structures in Fig. D.1 and in Fig. D.2 is that the Cu contact is replaced with a tunnel oxide to form a tunnel junction which is then used as the read port instead, whereas the tunnel junction on the bottom is used as the write port. The thicknesses of the oxide layers in the write port and in the read port of the DPMTJ proposed in [44] may be simultaneously made thinner and thicker, respectively, to reduce the resistance seen by I_{WRITE} and to increase the TMR of the read port. Thus, STT-MRAMs using the DPMTJ proposed in [44] can achieve better write-ability and readability than those using two-terminal MTJs.

D.2 The Domain-Wall MTJ Structure

The domain-wall based MTJ (DWMTJ) structure is another MTMTJ that has been proposed (Fig. D.3) [47]. The DWMTJ consists of a domain-wall stripe with complementary polarized pinned layers at the ends (i.e., the magnetization of the left PL is opposite that of the PL on the right as shown in Fig. D.3), and a free region between the pinned layers. A tunnel junction is formed on top of the free region, and the PL on top is used as the read port.

Write operations occur by passing I_{WRITE} between the PL's of the domain-wall stripe. Read operations occur by passing I_{READ} through the tunnel junction in the read port as shown in Fig. D.3. Note that the DWMTJ structure has all the advantages of the DPMTJ structure–separation of read and write current paths, low resistance in write current path to mitigate source degeneration, improved distinguishability and tunnel oxide reliability.

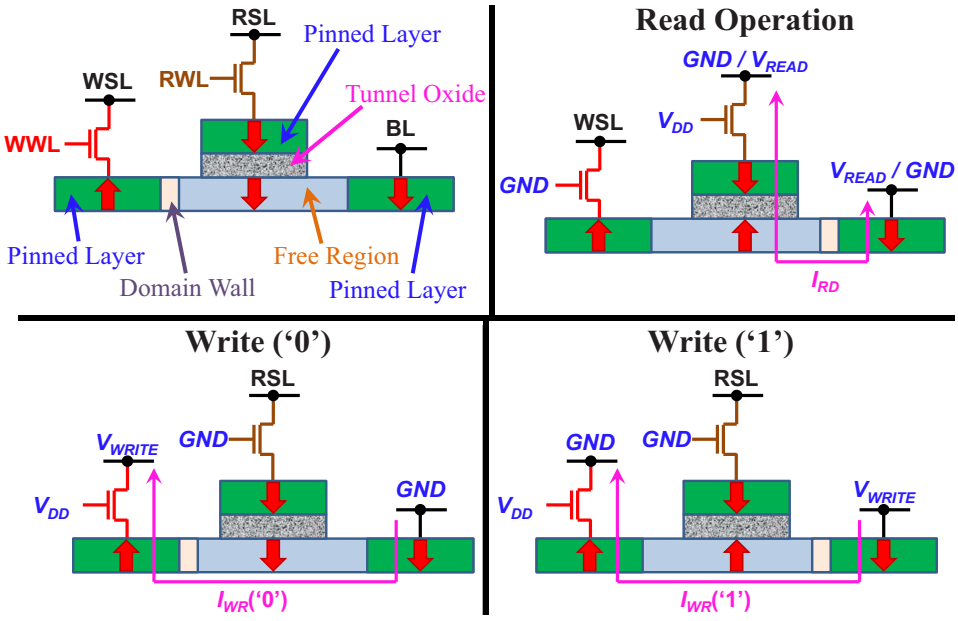


Fig. D.3. Structure of the domain-wall based MTJ (DWMTJ) proposed in [47]. I_{WRITE} flows in the domain-wall only whereas I_{READ} flows through the tunnel junction.

The biggest disadvantage of the DPMTJ and the DWMTJ structures is that the sensing scheme used in their read operation is single-ended in nature, just like in the STT-MRAM based on two-terminal MTJs. The reference, which may be generated separately, used for sensing the stored data in these STT-MRAM must be carefully chosen to optimize sensing failures in the presence of process variations. Self-referenced sensing schemes proposed in [30] and [31] improve the robustness of the single-ended sensing scheme and also eliminate the reference. However, the proposed sensing schemes require multiple sensing operations to correctly determine the stored data and degrade the read performance. Hence, self-referenced differential sensing schemes, which requires neither multiple sensing operations nor a separate reference, are desired to improve the read performance of STT-MRAM. To overcome this limitation, a novel MTMTJ structure that enables self-referenced differential sensing for read operations while preserving most advantages of DPMTJ and DWMTJ is proposed in Section 6.2.1 of this dissertation.

VITA

VITA

Xuanyao Fong (M06) received the B.Sc. degree in electrical engineering from Purdue University, West Lafayette, IN, in 2006, where he is currently working toward the Ph.D. degree in electrical and computer engineering. His research is focused on device-circuit-system co-design of spintronic systems with added emphasis on spintronic memory systems.

During January to August 2007, he was an Intern Engineer with Advanced Micro Devices, Inc., in the Boston Design Center, Boxboro, MA. He is currently a Research Assistant to Professor Kaushik Roy in the Nanoelectronics Research Laboratory, Purdue University. His research interests include device-circit-architecture co-design for Si and non-Si nanoelectronics and VLSI logic and memory systems using spintronic devices, circuits, and architectures. He will be continuing in the Nanoelectronics Research Laboratory in Purdue University as a Postdoctoral Researcher.

Mr. Fong received the AMD Design Excellence Award at Purdue in 2008, and the best paper award at the 2006 International Symposium on Low Power Electronics and Design.